Master's thesis
Theoretical and computational methods
Computational chemistry

Comparing descriptors for molecular clusters in unsupervised learning

Matias Jääskeläinen

May 28, 2020

Supervisors: Prof. Hanna Vehkamäki
PhD. Theo Kurtén

Examiners: Prof. Hanna Vehkamäki
PhD. Theo Kurtén

University of Helsinki
Faculty of Science
PL 64 (Gustaf Hällströmin katu 2a)
00014 Helsingin yliopisto

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Theoretical and computational methods Computational chemistry

Tekijä — Författare — Author

Matias Jääskeläinen

Työn nimi — Arbetets titel — Title

Comparing descriptors for molecular clusters in unsupervised learning

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	May 28, 2020	55

Tiivistelmä — Referat — Abstract

This thesis is about exploring descriptors for atmospheric molecular clusters. Descriptors are needed for applying machine learning methods for molecular systems. There is a collection of descriptors readily available in the `DScribe`-library developed in Aalto University for custom machine learning applications. The question of which descriptors to use is up to the user to decide. This study takes the first steps in integrating machine learning into existing procedure of configurational sampling that aims to find the optimal structure for any given molecular cluster of interest.

The structure selection step forms a bottleneck in the configurational sampling procedure. A new structure selection method presented in this study uses k-means clustering to find structures that are similar to each other. The clustering results can be used to discard redundant structures more effectively than before which leaves fewer structures to be calculated with more expensive computations. Altogether that speeds up the configurational sampling procedure. To aid the selection of suitable descriptor for this application, a comparison of four descriptors available in `DScribe` is made.

A procedure for structure selection by representing atmospheric clusters with descriptors and labeling them into groups with k-means was implemented. The performance of descriptors was compared with a custom score suitable for this application, and it was found that MBTR outperforms the other descriptors. This structure selection method will be utilized in the existing configurational sampling procedure for atmospheric molecular clusters but it is not restricted to that application.

Avainsanat — Nyckelord — Keywords

configurational sampling, machine learning, k-means, molecular descriptors

Säilytyspaikka — Förvaringsställe — Where deposited

eThesis

Muita tietoja — Övriga uppgifter — Additional information

Acknowledgements

I want to thank people who supported me during this spring. I am grateful to Theo Kurtén and Hanna Vehkamäki for their guidance, feedback and initial faith in this project when it was merely an idea. For Vitus Besel I want to tell my dearest gratitude for accepting me into their group on Introduction to Data Science -course, taking the first steps with me with this project and providing information, guidance and opinions whenever requested. I thank Jakub Kubečka, Anna Shcherbacheva and Marc Jäger for fruitful discussions and counselling on configurational sampling, mathematics, machine learning and descriptors. From you I have learned a lot during this spring.

Also Markus Rauhalahti, Patrick Rinke, Matthias Rupp, Nikolaj Tatti, Filippo Federici and Kai Puolamäki deserve recognition and my thanks for sharing their insight and expertise as I started to work on this topic. Furthermore I want to thank the Computational Atmospheric Physics group for their supportive scientific atmosphere that inspired me to take this research seriously.

My greatest gratitude goes to Sini Tenhunen for sharing this journey with me. With endless patience and marvellous cooking skills she gave me strength and faith to keep going even when my brains melted down and all seemed lost.

Contents

List of Symbols

Variables:

symbol	*explanation*	*unit*
E	Energy	kcal/mol
G	Gibbs free energy	kcal/mol
$\hat{H}$	Hamilton operator	
Ψ	Wavefunction	
ρ	Electron density	
R_g	Radius of gyration	Å
r	Spatial coordinates	m
τ	Spatial coordinates and electron spin	
$\hat{T}$	Kinetic energy operator	
$\hat{V}$	Potential energy operator	

List of Abbreviations

abbreviation	*explanation*
`ABCluster`	Artificial Bee Colony algorithm for cluster global optimization
`ACDC`	Atmospheric Cluster Dynamics Code
ACSF	Atom-Centered Symmetry Function
`ASE`	Atomic Simulation Environment
CC	Coupled Cluster
CI	Configuration Interaction
CM	Coulomb Matrix
CS	Configurational Sampling
DFT	Density Functional Theory
GFN-xTB	Geometry, Frequency, Noncovalent - eXtended Tight Binding
GM	Global Minimum
HF	Hartree-Fock
`JKCS`	Jammy Key for Configurational Sampling
LM	Local Minimum
MBTR	Many-Body Tensor Representation
ML	Machine Learning
MM	Molecular Mechanics
MSA	Methanesulfonic Acid
NPF	New Particle Formation
PES	Potential Energy Surface
SOAP	Smooth Overlap of Atomic Positions
SQM	Semiempirical Quantum Mechanics
t-SNE	t-Distributed Stochastic Neighbour Embedding
`XTB`	The program for GFN-xTB

1. Introduction

Configurational Sampling (CS) is a procedure for finding the optimal conformation for molecular systems ie. atoms bonded into molecules and groups of molecules held together by electronic interactions. Configuration, geometry and conformation refer to the arrangement of atoms and molecules in a molecular system. Systematic methods for finding the optimal conformation are actively studied. One quite recent method is a "build-up"-approach of the configurational sampling procedure presented by Kubečka et al. [Kubečka et al., 2019] and it is the method in the framework of this thesis.

First steps of the CS protocol includes a vast amount of calculations due to massive datasets of different configurations. Handling all the calculations requires a huge amount of computational power. Thus there is interest in reducing the amount of computation for example by discarding a set of redundant structures after each CS-step. The underlying motivation for this thesis is to study a method for making an intelligent automated decision to choose the collection of structures into the next CS-step. The collection should be as small as possible without losing the structure corresponding to the best geometry and energy.

Consequentially we tested an idea if similarities in a large set of atmospheric molecular clusters could be assessed with a Machine Learning (ML) method, since machine learning provides suitable algorithms for handling massive datasets. The first trials quickly showed that directly feeding a set of molecular structures into a clustering algorithm gives no meaningful results - at least not when the conventional .xyz-format is used. This led to research of different ways to represent molecular structures using a Python library called `DScribe` developed by Himanen et al. at Aalto University. `DScribe` provides tools for transforming a molecular structure into a representation – also called a *descriptor* – that a computer can use in various machine learning applications [Himanen et al., 2020].

In this study atmospheric molecular clusters are represented by four descriptors from the `DScribe`-library and the applicability of these descriptors in k-means clustering of atmospheric molecular clusters is compared. The k-means clustering is an unsupervised machine learning method for grouping together datapoints that are similar to each other. The goal in this study is to choose the best of four descriptors which enable the k-means clustering algorithm to group together molecular structures with similar geometries. A

structure selection procedure can then be executed based on the cluster assignments. Clustering structures represented by a descriptor is considered a good option for a structure selection method because it takes the whole structure into account when measuring similarities. It could work as a replacement or rather as an extension to current method which is representing the structures by their energy, dipole and radius of gyration values. In principle it is possible that two molecular clusters have similar dipole and radius of gyration values but different geometries. They can not however have similar descriptors provided that the descriptor is proper in terms of the descriptor criteria. Those criteria are covered in Chapter 5.1. With descriptors and a clustering based structure selection method it is also possible to select fewer structures from a large dataset than with the current structure selection method. Furthermore the aim is to implement this structure selection into a program that becomes part of the existing configurational sampling procedure and reduces computational costs. The hypothesis is that from the chosen four descriptors MBTR would stand out as the optimal choice. MBTR is by definition a global descriptor which is necessary when the variable of interest is a global variable like total energy. Moreover it is flexible and satisfies all criteria for a proper molecular descriptor.

Though machine learning on molecular systems has been done since *ab initio* quantum mechanics calculations started on computers [Rupp, 2015], the amount of publications has increased mainly during the 21st century and the last wave of ML development. One hindrance in the general utilisation of ML specifically in atmospheric chemistry research is that huge atmospherically relevant databases do not yet exist. One database of 633 atmospherically relevant molecular clusters is available [Elm, 2019] and could be used in research that utilises ML methods which can operate on smaller datasets. In contemporary quantum chemistry a typical example of applying machine learning is prediction of atomisation energies where descriptors are used quite routinely [Jung et al., 2020]. On the contrary clustering methods are used less frequently and the usage is concentrated on structure-activity research which focuses on finding molecules with similar properties [Lo et al., 2018, Holliday et al., 2004]. It appears that clustering methods have not been applied in configurational sampling proceedings which makes the work on this study particularly interesting. The methods proposed in this thesis are designed and applied for atmospherically relevant molecular systems, but they can be utilised in research of other molecular systems such as drugs or biomolecules.

Chapter 2 describes atmospheric new particle formation and the background for this study. Chapters 3 and 4 provide the necessary theory to understand computational quantum chemistry and the methods used in configurational sampling and machine learning. Chapter 5 introduces the descriptors that are used to represent the molecular clusters in this study. The methods used are presented in chapter 6. Finally the results, conclusions and suggestions for relevant future studies are presented in chapters 7 and 8.

2. Atmospheric new particle formation

In the field of atmospheric sciences there is a lot of interest in research of molecular clusters because they are an important factor in new particle formation (NPF) in the atmosphere. NPF affects the formation of clouds and clouds introduce effects that still remain highly unpredictable by current weather and climate models [Myhre et al., 2013]. For instance temperature and cloudiness are linked in a complex manner: temperature (and humidity) affect cloud formation and clouds affect the temperature by reflecting incoming shortwave radiation and by reflecting outgoing longwave radiation back to Earth's surface. A real life example: a cold winter night is forecasted to be overcast. If the model is wrong and the clouds do not appear, the net radiation from the Earth's surface will be higher which leads to the observed temperature being several degrees lower than the temperature forecast [Lentze, 2015]. At colder temperatures the clouds may form at lower levels than originally forecasted which affects aviation.

Figure 2.1 illustrates how trace gas molecules are connected with cloud formation. Trace gas molecules like sulfuric acid, water and ammonia molecules can form molecular clusters by binding through electronic interactions like hydrogen-bonding. Clusters grow further by condensation if the concentration of hydrogen-bonding molecules around them is high enough. Otherwise the molecules in the cluster will evaporate and the size of the cluster is reduced. After exceeding a critical radius of 1-3 nm a cluster will more likely grow further by condensation of trace gases and by coagulation with other clusters. In favourable conditions clusters grow further into cloud condensation nuclei (CCN) which can act as starting point to cloud formation.

Together with all the effects of which some are still unknown, the cluster growth to CCN size can take days [Kerminen et al., 2012]. The mechanisms for new particle formation remain still poorly understood. Some factors that affect the cluster formation are for example ion concentration which further depends on eg. lightning and cosmic rays [Svensmark et al., 2017]. Furthermore the cluster growth is affected by the environment like temperature and the purity of air [Kerminen et al., 2012]. Whether a cluster forms into a CCN that can start to form cloud droplets depends also on removal processes that

reduce cluster size or removes the cluster by coagulation to an existing CCN.

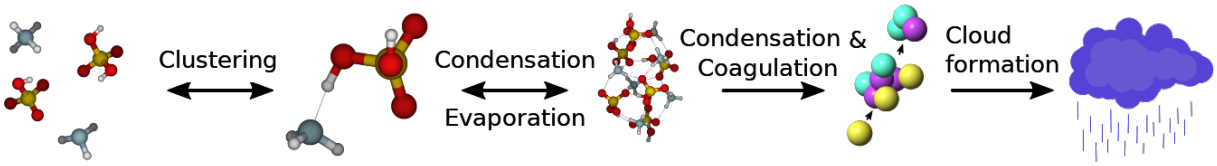


Figure 2.1: Trace gas molecules in the atmosphere form molecular clusters by binding through electronic interactions. Depending on the conditions the size of molecular clusters either grows by condensation or reduces through evaporation. After being formed the CCN affect cloud formation by acting as a surface for water vapour to start condensating into a droplet. Image courtesy of Vitus Besel, published with permission [Besel, 2020].

Studying the cluster kinetics by birth-death equations of molecular clusters with different amount of given gas molecules provides means to study first steps of atmospheric new particle formation. The birth-death equations and the new particle formation rate can be solved using a program called Atmospheric Cluster Dynamics Code (`ACDC`) by McGrath et al. [McGrath et al., 2012]. The accuracy of `ACDC` results depend largely on the quality of cluster structures and energies that are given as input. Therefore it is crucial to get as accurate conformations and energy values as possible in order to model the particle formation. Configurational sampling is used to obtain these accurate values. [Kubečka et al., 2019]

3. Theory of computational chemistry

The configurational sampling procedure includes calculations of molecular structures at different levels of theory. It starts with molecular mechanics -level and gradually moves towards more accurate *ab initio* -methods[†]. The desired outputs of those calculations are the energies and the optimized conformations of molecular clusters of given composition. The different calculation methods are presented here briefly with necessary literature. They are ordered by increasing level of theory which means also increasing accuracy and computational cost.

3.1 Molecular mechanics calculations

Molecular Mechanics (MM) is based on solving Newton's equations of motion for all atoms in the system. The atoms are modelled as points of masses that can have other properties like Van der Waals radius, charge and dipole. Bonds are modelled as springs between the masses mainly because the equation of motion for a spring is rather easy to solve. The parameters for bonds are equilibrium length and spring constant. Combined, the parameters form a so-called *force field*, which in total describes the interactions of the system [Jensen, 2017]. The parameter values for force fields can be obtained by fitting to quantum chemistry calculations done on an example system or they can be fitted empirically. There are also ready-made force fields like CHARMM [Brooks et al., 1983] and AMBER [Case et al., 2010] that are designed for certain type of systems.
The potential energy of the force field can be calculated for example as

$$\begin{aligned} V_{\mathrm{MM}} = &\sum_{i}^{N_{\mathrm{bonds}}} V_i^{\mathrm{bonds}} + \sum_{j}^{N_{\mathrm{angles}}} V_j^{\mathrm{angles}} + \sum_{k}^{N_{\mathrm{torsions}}} V_k^{\mathrm{torsions}} \\ &+ \sum_{i}^{N_{\mathrm{MM}}} \sum_{j>i}^{N_{\mathrm{MM}}} V_{ij}^{\mathrm{coulomb}} + \sum_{i}^{N_{\mathrm{MM}}} \sum_{j>i}^{N_{\mathrm{MM}}} V_{ij}^{\mathrm{LJ}} \end{aligned} \tag{3.1}$$

[†]Lat.: "from first principles"

where N_{MM} is the number of atoms, $V^{\mathrm{bonds}} = k_{bond}(r - r_0)^2$ is the bonding energy between atoms (k_{bond} is the force constant for the bond), $V^{\mathrm{angles}} = k_{angle}(\theta - \theta_0)^2$ (k_{angle} is the force constant for the bond) is the angle bending energy between three atoms, $V^{\mathrm{torsions}} = \frac{V_n}{2}[1 + \cos(n\varphi - \varphi_0)]$ is the energy of the rotational motion of bonds, $V_{ij}^{\mathrm{coulomb}} = \frac{q_i q_j}{\varepsilon r_{ij}}$ is the Coulombic force between atoms i and j, $V_{ij}^{\mathrm{LJ}} = \varepsilon\left[\left(\frac{r_{0_{ij}}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{0_{ij}}}{r_{ij}}\right)^{6}\right]$ is the Lennard-Jones potential that models Van der Waals interactions [Jensen, 2017].

The molecular mechanics calculations are so fast that they can be used in the genetic algorithm that conducts the exploration of different configurations in the first step of configurational sampling.

3.2 wavefunction methods

In quantum mechanics, the system is *described* by a wavefunction Ψ. The properties for the system can be obtained by solving the time independent Schrödinger equation:

$$\widehat{\mathrm{H}}\Psi = E\Psi \tag{3.2}$$

where $\widehat{\mathrm{H}}$ is the Hamilton operator, Ψ is the wavefunction, E is the energy of the system.

Equation 3.2 is an eigenvalue equation which means that there is an operator (here $\widehat{\mathrm{H}}$) that operates on the eigenfunction (here Ψ). The operation yields an eigenvalue (here E) and the same eigenfunction. In quantum chemistry the operator corresponds to a physical quantity and the operation corresponds to a measurement that yields the value.

The Hamilton operator corresponds to the energy of the system by defining the interactions of the system. Ignoring the relativistic effects the molecular Hamilton operator becomes:

$$\widehat{\mathrm{H}} = \widehat{\mathrm{T}}_e + \widehat{\mathrm{T}}_N + \widehat{\mathrm{V}}_{ee} + \widehat{\mathrm{V}}_{eN} + \widehat{\mathrm{V}}_{NN} \tag{3.3}$$

where $\widehat{\mathrm{T}}_e$ is the kinetic energy of the electrons, $\widehat{\mathrm{T}}_N$ is the kinetic energy of the nuclei, $\widehat{\mathrm{V}}_{ee}$ is the potential energy operator for the electron-electron interactions, $\widehat{\mathrm{V}}_{eN}$ is the potential energy operator for the electron-nucleus interactions, $\widehat{\mathrm{V}}_{NN}$ is the potential energy operator for the nucleus-nucleus interactions.

The first approximation made when solving molecular Schrödinger equations is the Born-Oppenheimer approximation [Born and Oppenheimer, 1927] which relies on the fact that the atom nuclei are three orders of magnitude heavier than electrons. Thus the movement of nuclei can be separated from the movement of the electrons.

The wavefunction Ψ is the solution for the Schrödinger equation. The true form of Ψ is the perfect representation of given system, because all the properties of the electronic ground state of the system could be calculated simply by operating on Ψ with a quantum

mechanical operator [Rupp, 2015]. The true form of Ψ is not known except for the most simple systems like a free particle, or a hydrogen atom. For systems more complex than molecular hydrogen ion (H_2^+) the Schrödinger equation is not analytically solvable but requires approximative methods.

Generally, in *ab initio* wavefunction methods the approximated wavefunction of the molecular system is made by forming a linear combination of Gaussian functions. Other functions can be used for different systems: for example large molecular surfaces benefit from using sinusoidal functions. With proper coefficients the functions form a basis set that functions as an approximation to the wavefunction of the system. The size of the basis set - namely how many gaussian functions are used - is one factor determining the accuracy of the calculation.

The basic method of approximating the Schrödinger equation is the *Hartree-Fock* (HF) method [Hartree, 1928, Fock, 1930]. For a molecular cluster with approximately 20 atoms one HF calculation takes some seconds or minutes even on a standard laptop computer. With a large basis set HF methods can yield about 99% of the total energy which mainly includes contribution from the chemically irrelevant core electrons [Jensen, 2017]. The remaining one percent results from the electron-electron correlation and it contains most of the contributions from the chemical bonding, which Hartree-Fock method largely neglects. HF treats electrons as they were moving in the mean electromagnetic field of other electrons. In reality moving electrons cause far more complicated interactions to each other - namely the electron correlation which can then be taken into account with more sophisticated methods.

There are three ways to improve Hartree-Fock. It can be parametrised which yields the semiempirical methods that can calculate faster – but the accuracy depends on the parameter values. The very problem setting can be reformulated calculating the energy from the electron density ρ instead of the wavefunction Ψ which then yields basic DFT (see Chapter 3.2.2). Additionally it can be made more accurate by calculating the electron correlation which makes the calculation computationally expensive.

3.2.1 Semiempirical methods

Semiempirical Quantum Mechanical (SQM) methods approximate HF or DFT calculations by parametrisation of the equations. It leads to omitting some or all of the differential overlaps between atomic basis functions. This saves computation time because there are fewer one- and two-electron integrals left to calculate. The parameters for the semiempirical equations are obtained from the results of reference data generated by hybrid DFT calculations - a procedure which is similar to the fitting of force field parameters in MM simulations. This fitting of parameters into equations in order to obtain more ac-

curate results is also comparable to machine learning (Chapter 4) and regression methods. There the coefficients of the model are fitted with training data and the accuracy and applicability of the model then depend on the quality and features of the training data.

In this thesis a semiempirical method called GFN-xTB is used in calculation of molecular cluster energies. GFN-xTB stands for Geometry, Frequency, Noncovalent - eXtended Tight Binding method. It is designed to yield reasonable accuracy in geometries, vibrational frequencies, and noncovalent interactions. GFN-xTB provides higher accuracy for the target properties than existing 'general-purpose' semiempirical approaches for molecules with atoms from the whole periodic table. [Grimme et al., 2017] GFN-xTB is implemented in a program called `XTB`. For a molecular cluster with approximately 20 atoms one `XTB` calculation with geometry optimisation takes a few seconds and hence it is feasible to use for all structures in the first steps of the configurational sampling procedure. The accuracy of `XTB` is not enough for calculating Gibbs free energies for atmospheric molecular cluster formation studies.

3.2.2 Density Functional Theory

Density Functional Theory (DFT) is a method of calculating electronic energies from the electron density ρ instead of the wavefunction Ψ. In DFT the system is described by using the electron density ρ which is connected to the electronic wavefunction as

$$\rho = N \int |\Psi(\tau_1, \cdots, \tau_N)|^2 d\tau_1, \cdots, d\tau_N \tag{3.4}$$

where τ is the combination of spatial coordinates r of an electron and its spin, and N is the number of electrons

It is proven by Hohenberg and Kohn [Hohenberg and Kohn, 1964] that the external potential $\vartheta(r)$, and hence the total energy of a system, is a unique functional of the electron density $\rho(r)$ and that the density which minimizes the total energy is the exact ground state density. Thus DFT energies can be calculated using the variational principle - the same used with HF calculations.

The DFT energy, taking Born-Oppenheimer approximation [Born and Oppenheimer, 1927] into account can be written as

$$E_{\mathrm{DFT}} = E_T + E_{eN} + E_J + E_K + E_C \tag{3.5}$$

where E_J is the Coulomb interaction of the electrons, E_{eN} is the electron–nucleus interaction, E_T is the kinetic energy of the electrons, E_K is the exchange term and E_C is the correlation term.

The first two can be calculated as Coulomb interactions, but for the last three terms an approximation by Kohn and Sham [Kohn and Sham, 1965] is commonly used. The

Kohn-Sham energy E_{KS} is

$$E_{KS}[\rho] = E_{TS}[\rho] + E_J[\rho] + E_{eN}[\rho] + E_{XC}[\rho] \tag{3.6}$$

where the exchange-correlation part $E_{XC}[\rho]$ is

$$E_{XC}[\rho] = (E_T^{\text{exact}}[\rho] - E_{TS}[\rho]) + (E_{ee}[\rho] - E_J[\rho]) \tag{3.7}$$

The terms in $E_{XC}[\rho]$ are those that are unknown and different functionals have been developed to approximate them. Different levels of functionals are described by Jacob's ladders [Perdew and Schmidt, 2001], in which higher level functionals are more accurate and require more computational resources. DFT can also be a part of a hybrid method that uses exchange energies calculated with HF combined with correlation energies from DFT methods. For a molecular cluster with approximately 20 atoms one DFT calculation with geometry optimisation can take a few hours and hence it is preferred to minimise the amount of structures in the DFT calculation step of the configurational sampling procedure. DFT often yields results with errors less than 2 kcal/mol [Koch and Holthausen, 2001] but higher accuracy is necessary because many variables of interest like equilibrium constants depend exponentially on Gibbs free energies that DFT frequency calculations yield.

3.2.3 wavefunction methods with electron correlation

As mentioned in Chapter 3.2 basic HF method can yield at maximum 99% of the total energy of any molecular system. This is due to the approximations made on electronic interactions that completely ignore the electron correlation caused by constantly changing interaction between moving electrons. Multiple electron correlation methods such as Configuration Interaction (CI) and Coupled Cluster (CC) have been developed for calculating the remaining 1% of the energy because it is often important for describing chemical phenomena eg. chemical bonding. [Jensen, 2017]

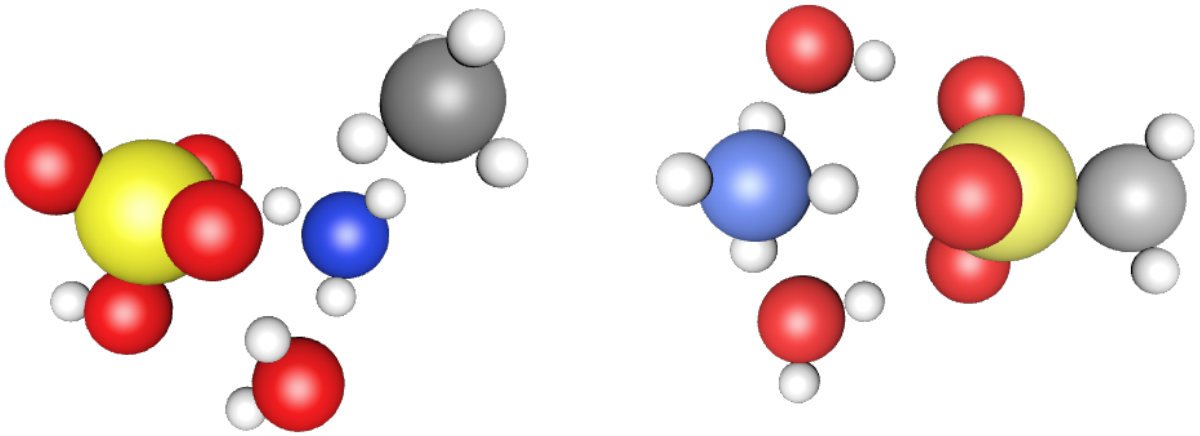
The remaining 1% is also crucial for obtaining accurate Gibbs free energies of atmospheric molecular clusters for modelling the cluster growth. For a molecular cluster with approximately 20 atoms single point energy calculations with CCSD(T)* can already take a few days. Thus the purpose of configurational sampling procedure is to result only a few structures that enter the most accurate computation step.

*A Coupled Cluster calculation with Single, Double and perturbative Triple excitations - a "gold standard" in computational chemistry

3.3 Finding the optimal molecular geometry

Both conformation and geometry of a molecular structure refer to the relative positions of the atoms in a molecule or a molecular cluster. Molecular structures are called "conformers" if they have same atoms and same bonding pattern between the atoms but different geometries. The geometry of the structure changes as the atoms move which happens constantly: the bonds vibrate, functional groups and whole molecules rotate in relation to other parts of the structure.

The term configuration in "configurational sampling" is broader than conformation - comparable to a general term "structure". If two molecular clusters have different conformations they have same atoms but the ordering of atoms is not restricted. Figure 3.1 shows two molecular clusters with same atoms, but with different molecules: in Fig. 3.1(a) carbon and sulphur atoms are in different molecules but in 3.1(b) they are in the same molecule. Therefore the structures are not called conformers but conformations of each other.



(a) A molecular cluster of methane, sulfuric acid, water and ammonia.

(b) A molecular cluster of methanesulfonic acid, water and ammonia

Figure 3.1: Two different configurations illustrated with molecular cluster that have same atoms, but the molecules are different. The figure shows sulphur in yellow, oxygen in red, nitrogen in blue, carbon in grey and hydrogen in white.

Finding the accurate geometry of a molecular system is always a balance between the computational cost, calculation time and the desired accuracy. Current *ab initio* quantum mechanics methods for molecular calculations quickly become computationally expensive as the size of the system grows [Huo and Rupp, 2017]. On the other hand molecular mechanics methods based on classical physics can handle larger systems but

are not accurate due to the approximations made.

The accurate modelling of structures of any given system is important in all computational molecular research, because the properties of the molecular system depend largely on the structure. In atmospheric molecular cluster research (described in Chapter 2) one property in focus is the Gibbs free energy which is obtained from the vibration frequency calculations. The value of Gibbs free energy is related to the stability of a molecular cluster: when a molecular cluster is formed the change in Gibbs free energy is proportional to the equilibrium constant of the formation reaction. The average value for Gibbs free energy of a molecular structure can be obtained as a sum over all conformations i of the given structure:

$$G = -kT \ln \sum_{i=0} e^{-\frac{G_i}{kT}} \tag{3.8}$$

where k is the Bolzmann constant, T is the temperature and G_i is the Gibbs free energy for one conformer i.

Gibbs free energy values for a conformer can be obtained from *ab initio* calculations like DFT and CC, but the accuracy needed for eg. `ACDC` requires the use of CC. Even with a low level of theory calculating all conformers through is unfeasible. Due to the fact that the value of G in Eq. 3.8 is proportional to $e^{\frac{-G_i}{kT}}$ the approximation is made that it is enough to compute the Gibbs free energy of the conformer with lowest Gibbs free energy value obtained from the configurational sampling procedure. The approximation is valid if there actually is one conformer with significantly lower value of G_i, but the bigger molecular cluster is the more likely there will be multiple conformers with similar G_i values. With recent increase of computational power and the development of systematic configurational sampling methods (in which also this thesis is contributing) it starts to be possible to account for all low free energy conformers that have a significant contribution to the average Gibbs free energy value. [Partanen et al., 2016]

3.3.1 The Potential Energy Surface

The Born-Oppenheimer approximation [Born and Oppenheimer, 1927] states that the electronic wavefunction of a molecule can be solved separately from the nuclear wavefunction. Thus the nuclei move in the potential created by the electronic interactions. Calculating the electronic energy for all possible locations of the nuclei yields a multidimensional Potential Energy Surface (PES). As shown in Fig. 3.2 the geometry of a molecule is directly mapped to the PES. The values of PES gives insight of the stability of the structure: lower potential energy value imply stability.

The dimensions in PES corresponds to the degrees of freedom that the system has. With molecular structures the degrees of freedom come from all movement that the atoms

in the system can have in relation to other atoms namely bond stretches, angle bendings and dihedral twists. Nonlinear molecules have $3n - 6$ degrees of freedom where n is the number of atoms in the system. The substracted six degrees of freedom correspond to the rotation and movement of the system as a whole. The approximation of rigid intramolecular bonds changes the amount of degrees of freedom to $3M - 6$ where M is the amount of molecules in the system. In Equation 3.1 this means omitting the terms corresponding to bonds, angles and torsions. This effectively reduces the dimensions of the PES which reduces the amount of necessary computation.

Optimising a structure with conventional wavefunction and molecular mechanics methods can be pictured in two dimensional PES as "sliding down" the potential energy surface until a minimum is found. That inherently means that the global minimum of PES can be found only by starting from a point of the surface that lies "uphill" from the GM, but not from a point that lies behind a peak. [Jensen, 2017] In the Figure 3.2 this is illustrated with two dashed arrows starting from slightly different sides of a peak and ending in two different minima. In molecular structure calculations this means that the starting geometry has to be very close to the geometry that maps to the GM which is very difficult with complex structures that molecular clusters can have.

The need of overcoming this dependency of the initial structure has led to the development of methods to find the most stable conformation of any given molecular system by modelling the PES with a sufficient accuracy like basin hopping, umbrella sampling, simulated annealing and genetic algorithms. In this thesis the focus is in the build-up approach configurational sampling procedure presented by Kubečka et al. [Kubečka et al., 2019].

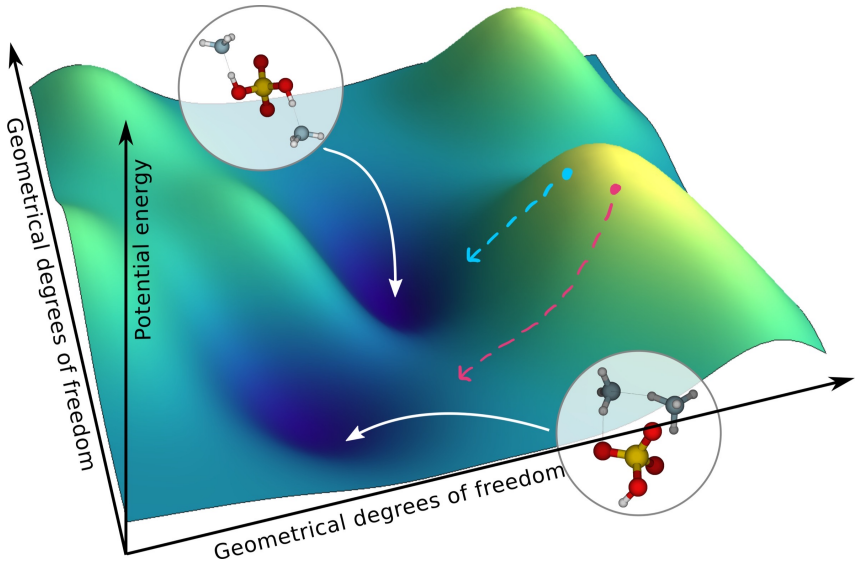


Figure 3.2: A potential energy surface (PES) of a molecular structure visualised in two dimensions. The axis correspond to the geometrical degrees of freedom, which can be eg. bond stretching. Each point on a PES matches to a set of atom positions: the distance between atoms define the strength of their interaction which defines their potential energy. Peaks of the PES coincide with structures that have high energy and hence are unstable. Accordingly the valleys in the PES - called local minima - fall in with structures that have low energies and are more likely to be stable. The lowest point of the surface is called the global minimum. Each degree of freedom has an own dimension in a real PES of a molecular system. When visualising potential energy surfaces only two degrees of freedom can be visualised at a time: the third dimension is the energy value. Image courtesy of Vitus Besel, published and edited with permission [Besel, 2020].

3.3.2 Artificial Bee Colony algorithm

Artificial Bee Colony algorithm is a genetic algorithm that is designed for molecular structure sampling with the aim to construct a large amount of different conformations [Zhang and Dolg, 2015, Zhang and Dolg, 2016]. The algorithm is implemented in `ABCluster` program which takes the optimised structures of the constituent molecules as inputs and gives different cluster geometries as output. The amount of computation can be reduced by fixing the geometries of input molecules. Effectively that reduces the dimensions of the PES from $3n-6$ into $3M-6$ where $M < n$. Then all the variation between the output clusters comes from the positions of molecules in a cluster. [Kubečka et al., 2019]

`ABCluster` tries to mime the behaviour of honey bee colonies in search of the best food source. The algorithm creates a set of random molecular clusters as initial guess and performs an exploration on potential energy surface according to the random sample. Good structures are saved and investigated whether they have changed from previous rounds. The energy of the cluster in molecular mechanics level works as the quality measure of the cluster. If a structure did not change for a specified number of cycles they are saved, new trials are created and the algorithm starts its next cycle. The assumption

is that `ABCluster` delivers a sufficiently large sample that covers the local minima of the MM-level potential energy surface including the global minimum. [Zhang and Dolg, 2015, Zhang and Dolg, 2016, Kubečka et al., 2019]

3.3.3 Configurational sampling

A systematic configurational sampling (CS) procedure to find the global minimum (GM) of a potential energy surface (PES) is implemented in a program `JKCS` (Jammy Key for Configurational Sampling) by Kubečka et al. [Kubečka et al., 2019] and illustrated in Figure 3.3. The procedure starts with `ABCluster` exhaustively sampling the PES on a molecular mechanics level and saving a chosen amount of local minima for further computation. GFN-xTB is then used for re-optimisation and calculation of energy values at semi-empirical level of theory. The structures output from XTB are sequentially filtered and calculated with more and more accurate wavefunction methods until only a handful of structures remain. Effectively the structures leaving each calculation step correspond to the local minima of the PES on given level of theory. The last step is to calculate properties like Gibbs free energy for the remaining few structures and use those values for example as an input in `ACDC` for cluster birth-death equation calculations.

The structure selection steps are currently done by the values of energy, dipole and radius of gyration of the molecular cluster. A dipole of a system is caused by different electronegativities of the atoms leading to uneven distribution of the electronic density. The radius of gyration R_g is proportional to the moment of inertia of a rotating system and it can be calculated from:

$$R_g^2 = \frac{\sum_{i=1}^{N} m_i |\bar{r}_i - \bar{r}_{COM}|^2}{\sum_{i=1}^{N} m_i} \tag{3.9}$$

where N is the number of atoms in the molecular system, m_i is the mass of atom i, $\bar{r}_i$ the position of atom i and $\bar{r}_{COM}$ is the position of the center of mass of the molecular system.

This thesis proposes a new structure selection method to be used instead or in addition to the existing method. The proposed method uses machine learning to aid the selection because machine learning algorithms are designed for handling large datasets. Instead of dipole and radius of gyration the proposed method uses predesigned descriptors to represent molecular clusters to a clustering algorithm that labels similar molecular systems to corresponding clusters to make the selection more effective.

The systems studied with CS procedure can be in general any molecules or clusters of molecules, but in this thesis the focus is on one cluster at a time so also term "conformation sampling" could be used. To keep consistent with the literature the term “configurational sampling” is used.

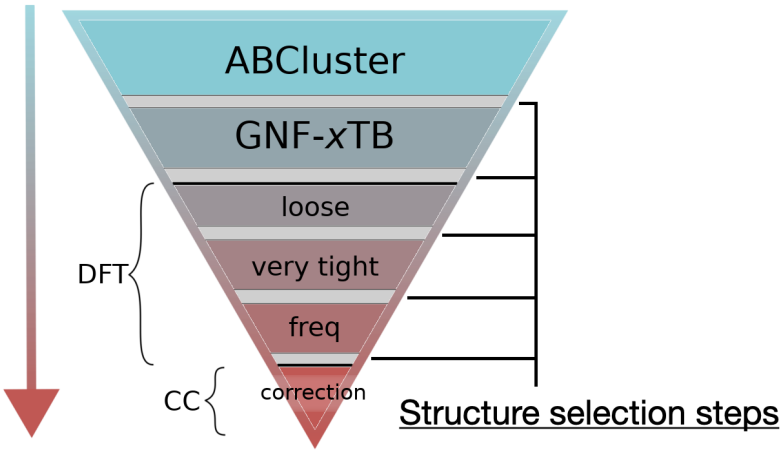


Figure 3.3: Illustration showing the procedure of configurational sampling as implemented in `JKCS` program. The width of the triangle symbolizes the amount of molecular cluster structures in each step of the procedure. Each step consist of energy and optimisation calculations followed by structure selection steps where redundant structures are discarded. The steps loose and very tight refer to geometry optimisation convergence criteria and freq refers to vibration calculation. The final output is a handful of structures that are used in further research eg. in `ACDC`. Image courtesy of Vitus Besel, edited and published with permission [Besel, 2020]

4. Methods of machine learning

Machine Learning (ML) has been around already from the 50's. It has gone through hypes and downtrends but over the past two decades it has developed drastically mainly due to increase in computational power, the development of new learning algorithms and the availability of massive datasets. Together they have made many interesting applications possible, for example computer vision, natural language processing and speech recognition [Jordan and Mitchell, 2015]. One thing in common with all of these applications is that they require handling of vast amount of data.

At the time of writing in 2020 a Google image search with keywords "machine learning" gives mainly blueish pictures where a human head has some network in it. This is probably due to a field of machine learning called Neural Networks (NN) but ML covers much wider collection of methods than just NN. Machine learning can be described as a study of computer algorithms to build systems that automatically improve through experience. Main characteristics of ML algorithms is that they operate on massive datasets utilizing the methods of mathematics, statistics and computer science. The algorithms use the data to construct a model that allows the computer to operate on a given task. A computational chemistry example of a machine learning application would be that instead of calculating DFT on every new system that appears, one can use a finite number of existing DFT results to train a machine learning model. The training here means minimizing a loss function of the model between the atomistic structure and its properties. That can be thought as analogous to the structure-property relation in quantum mechanics [Himanen et al., 2020]. The computer can use the model to predict DFT energies for new systems. Such prediction is already possible but it works only on similar molecules than the ones used in model training. For example adding a new element to the system in study would introduce features (eg. interactions) that the model does not know how to take into account since the new element was not part of the training set. That leads to larger uncertainty in the prediction depending on the model and method in use.

In computational chemistry machine learning has been applied for example in predicting molecular properties such as orbital energies, atomization energies [Stuke et al., 2019, Rupp et al., 2012, Jung et al., 2020] and energy predictions for solids

[Seko et al., 2017]. Also neural networks have already seen the daylight in solving quantum chemical problems [Smith et al., 2019a, Gebauer et al., 2018, Ghosh et al., 2019, Schütt et al., 2019b, Yao et al., 2018, Schütt et al., 2019a]. The advantage in machine learning is that instead of solving the Schrödinger equation 3.2 the solutions can be statistically estimated by training a model based on a reference set of known solutions. That will significantly reduce the computational costs by skipping the redundant work done in calculating quantum mechanical calculations for similar systems with correlating output [Huo and Rupp, 2017, Rupp, 2015, Ramakrishnan et al., 2015, Himanen et al., 2020].

4.1 Types of learning

Machine learning techniques are divided into subfields as shown in the Figure 4.1. Two most common ones are supervised learning and unsupervised learning. They are used in different tasks like clustering and regression. Tasks are suitable for various applications: eg. clustering is regularly used in recommender systems and customer segmentation.

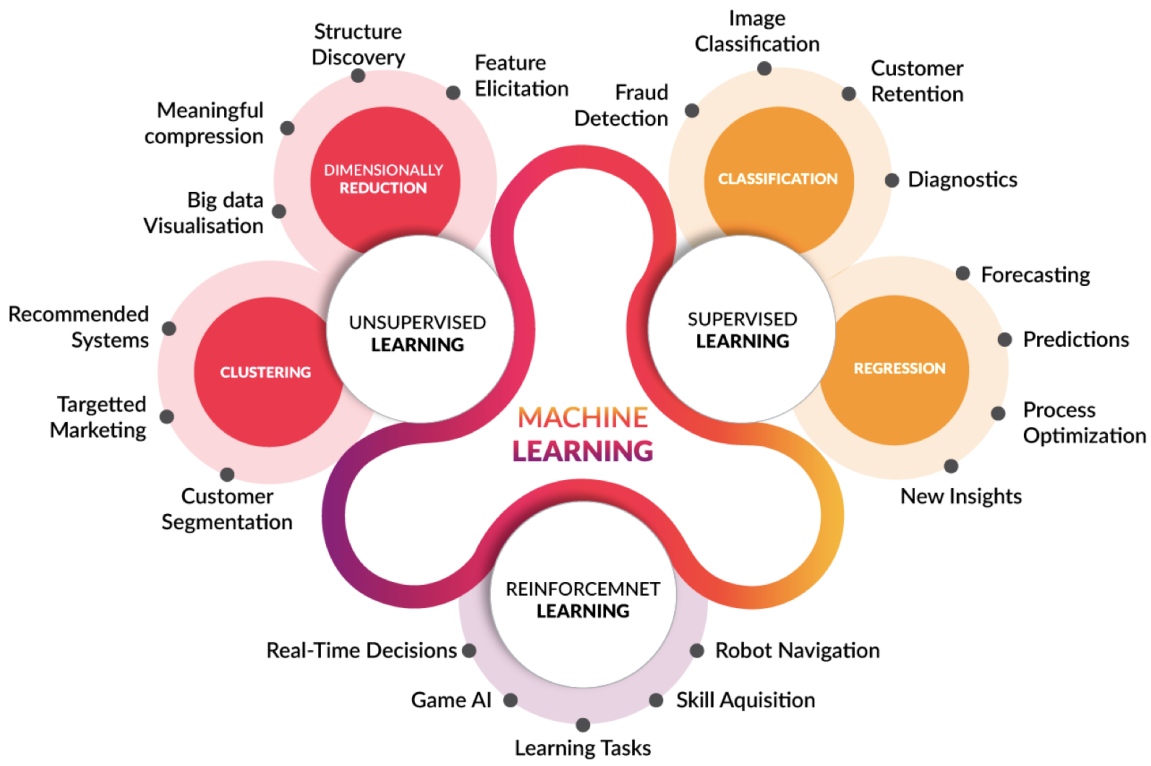


Figure 4.1: From the edges to the middle: Machine learning has many applications that are grouped by the tasks that are categorized under subfields of ML. Image from [Heidenreich, 2018].

In machine learning problem it is common to categorize the data as *feature* variables and target variables or *labels*. Using the Iris dataset [Dua and Graff, 2017] as a textbook example: the lengths and widths of the flower petals are used as *features* and the flower

species are used as *labels*. Whether to use supervised or unsupervised learning depends obviously on the problem at hand, but also on the data available. Supervised learning methods need both features and labels for the model training, but unsupervised learning methods are designed to operate on features only.

Supervised learning methods train a model that maps *features* into *labels* according to the training data. The model is used to predict the labels of new yet unknown data. Unsupervised learning methods are used when the true *labels* are not known or available. They use multiple *feature* variables to find eg. patterns or similarities which is useful for early stage exploratory data analysis or finding groups of similar molecular clusters as this thesis aims to demonstrate.

4.2 K-means clustering

"Clustering looks to find homogeneous subgroups among the observations" [James et al., 2017]. K-means is a clustering method that can be used to group datapoints together in a way that points in one k-means cluster are more similar to each other than to points in other clusters. Points of data are clustered into clusters by minimizing the *total intra-cluster variation* (or *total within-cluster variation*) defined as:

$$\sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{4.1}$$

where x_i is a data point in a cluster C_k, μ_k is the mean value of the points in cluster C_k, k is the number of clusters.
This is the summation of all the clusters over the sum of squared distances between items and their corresponding centroid. The notion of similarity is then derived by how close a data point is to the centroid of the cluster. The measure of distance, also called "similarity measure" is a metric that reflects the strength of relationship between two data objects. Equation 4.1 is using Euclidian distance as a similarity measure.

The algorithm for calculating k-means is introduced by MacQueen (1967) and the steps are illustrated in Figure 4.2. The process starts with randomly adding K cluster centroids (i.e. *vectors of means*) and assigning data to clusters based on the distance from the centroids. The next step is to update cluster centroids into the centers of their respective clusters by feature averages of the objects in each cluster respectively. The centroids are moved accordingly into the "middle" of the clusters and the algorithm continues iteratively until nothing changes in the centroids or the cluster assignments. [James et al., 2017]

K-means is used for exploratory data mining, pattern recognition, image analysis, information retrieval, bio-informatics, data compression, computer graphics and

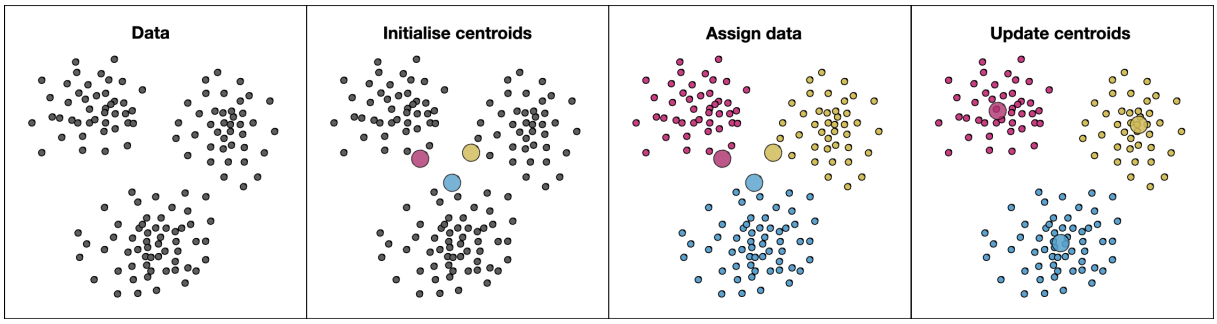


Figure 4.2: The algorithm for applying k-means clustering on toy data. The algorithm loops iteratively with the last two images until convergence. Each step the datapoints are assigned to clusters according to their distance from the cluster centroid and the centroids are updated into the middle of their respective clusters.

on this thesis for the structure selection step in configurational sampling procedure. [Raykov et al., 2016, James et al., 2017, Wei Zhong et al., 2005]

The centroid initialisation of k-means algorithm is stochastic which means that every time the algorithm runs the results may differ. The initialisation can be made deterministic by providing a value for the random seed before running the algorithm. The k-means implementation in `sklearn` features a more intelligent stochastic initialisation method by Arthur and Vassilvitskii which make the algorithm converge faster than with using the original initialisation method [Arthur and Vassilvitskii, 2007]. In this study that stochasticity is exploited in descriptor comparison as described in Chapter 6.6.

5. Molecular representations

The success of predicting molecular energies accurately relies on how the molecules are represented to the machine learning algorithm. There are many ways to store molecule structure information into a computer. Most common would be the XYZ-format where each atom of the molecule has a label and cartesian coordinates. Many computational chemistry software also understand so called Z-matrix where the position of each atom is defined in relation to the neighboring atoms. Those formats are not suitable for the majority of the machine learning algorithms. [Huo and Rupp, 2017, Himanen et al., 2020]

The machine learning algorithms operate on features of the data points - here molecular structures. Those features must behave rigorously with respect to the variable of interest - here energy. Usually the machine learning applications on molecular systems require some *feature engineering* to yield meaningful results. The features selected in feature engineering - hereby referred as a *descriptor* - encode the chemical identity of the molecular system ie. chemical composition and atomic configuration into a form that is interpretable for the machine learning algorithm without confusion. Therefore descriptors are a crucial ingredient for the development of machine learning models for atomistic systems. The connection between the descriptor and the properties of the system can be thought as analogous to the connection between the properties of the molecule and the wavefunction of that molecule. [Von Lilienfeld et al., 2015, Rupp, 2015]

5.1 Features of molecular descriptors

Here is a concise list of most important criteria that a good descriptor should fulfill in order to make reliable machine learning possible.

1. Invariance with respect to labelling and numbering of the atoms in the molecule:
    - The value of a descriptor must not depend on how the molecule atoms are labelled or numbered.
2. Invariance with respect to the molecule rotation and translation:

- The value of a descriptor must not depend on the absolute values of numerical coordinates defining the atom positions with respect to some arbitrary origin.

3. A definition which is unambiguous and algorithmically computable:

   - A molecular descriptor must be defined by a computable mathematical expression whose terms have to be unambiguous and clearly defined by the molecular structure. For example a Coulomb matrix element is defined by the Coulombic force between an atom pair.

4. The values in a suitable numerical range for the set of molecules where the descriptor is applicable to:

   - The values of a molecular descriptors must be in an acceptable numerical range. For example, descriptors defined on the product of some atomic property quickly reach large numerical values for big molecules.

5. The descriptor must have a structural interpretation and it has to be unique:

   - The structure of the molecular system can be decoded from the descriptor, and no other structures correspond to the same descriptor values. Similarly a molecular structure can be encoded to only one set of descriptor values. For example two stereoisomers that have mirrored geometries and hence same forces between the atom pairs result similar ordinary Coulomb matrices. Thus an ordinary Coulomb matrix is not an ideal descriptor although it is used a lot for its interpretability.

6. The descriptor should correlate with at least one property of a molecule.

   - For example the values of a Coulomb matrix correlate with the distances between the atoms in a molecule. That leads to some degree of correlation with the energy of the molecule.

7. The values of a descriptor should be continuous:

   - A gradual change in descriptor values should correspond with gradual change in the molecular structure.

8. The definition of a descriptor should not include experimental properties.
9. A Descriptor should not be restricted to a too small class of molecules.

The list is collected from multiple sources. [Himanen et al., 2020, Behler, 2011, Von Lilienfeld et al., 2015, Bartók et al., 2013, Rupp, 2015, Huo and Rupp, 2017].

5.1.1 Local and global descriptors

Descriptors can be divided into global and local descriptors according to the way that they are constructed. Global descriptors encode the whole structure of the molecule and are suitable for predicting the values for global properties like molecular energies, formation energies and band gaps. Global descriptors are conceptually like the wavefunction of the system - unlike local descriptors. Moreover their computation may not scale linearly with the size of the system. Local descriptors encode the chemical environment of each of the atoms individually and are suitable for predicting local properties like atomic forces, adsorption energies, or properties that can be summed from local contributions. Local properties depend on the immediate chemical environment of each atom. Transforming a local descriptor to a global one can be done by simply averaging the outputs of multiple local sites, developing a custom kernel to combine information from multiple sites, or the predicted property can in some cases be directly modelled as a sum of local contributions. [De et al., 2016, Himanen et al., 2020, Jung et al., 2020]

5.2 Descriptors in this study

In this study the interests lies within the conformations and the potential energy surface of a given atmospheric molecular cluster. Thus the descriptors have to be able to encode the clusters' 3D-structure including changes in the conformation. There are many descriptors available that suit the problem at hand [Bartók et al., 2013, Behler, 2011, Rupp et al., 2012, Huo and Rupp, 2017, Hansen et al., 2015, Faber et al., 2018]. The descriptors chosen for this study are all included in a Python library called `DScribe` [Himanen et al., 2020] which provides a concise way of using all implemented descriptors in custom machine learning applications. The set of descriptors used in this study is presented in table 5.1.

Table 5.1: The descriptors used in this study can be divided to global or local descriptors. Applications that need mapping between total energy and the structure need global descriptors. Local descriptors are made global by averaging their values over all atoms.

Desc	Abbreviation	Global	Local
Coulomb Matrix	CM	x	
Many-Body Tensor Representation	MBTR	x	
Atom-centered Symmetry Functions	ACSF		x
Smooth Overlap of Atomic Positions	SOAP		x

5.2.1 Coulomb Matrix

A Coulomb Matrix (CM) [Rupp et al., 2012, Montavon et al., 2015] represents a molecular system by a matrix of coulomb forces between each pair of atoms:

$$\boldsymbol{M}_{ij}^{\text{Coulomb}} = \begin{cases} 0.5Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{R_{ij}} & \text{for } i \neq j \end{cases} \tag{5.1}$$

where Z is the atomic charge of atoms i and j and R_{ij} is the distance between atoms i and j.

For the charges of the atoms types standard atomic charges are used. Figure 5.1 shows a molecular cluster used in this study and the respective Coulomb matrix calculated with `DScribe`. A Coulomb matrix takes into account the charges and the distances, so it cannot for example distinguish between enantiomers [Ramakrishnan et al., 2015]. The values of CM change when the stoichiometry (no. of atoms) or configuration changes. As such CM:s are not invariant to the permutations of the atomic indices, but the invariance is accomplished by sorting them by their Euclidian norm [Montavon et al., 2015]. The `DScribe`-package uses sorted CM:s as the default.

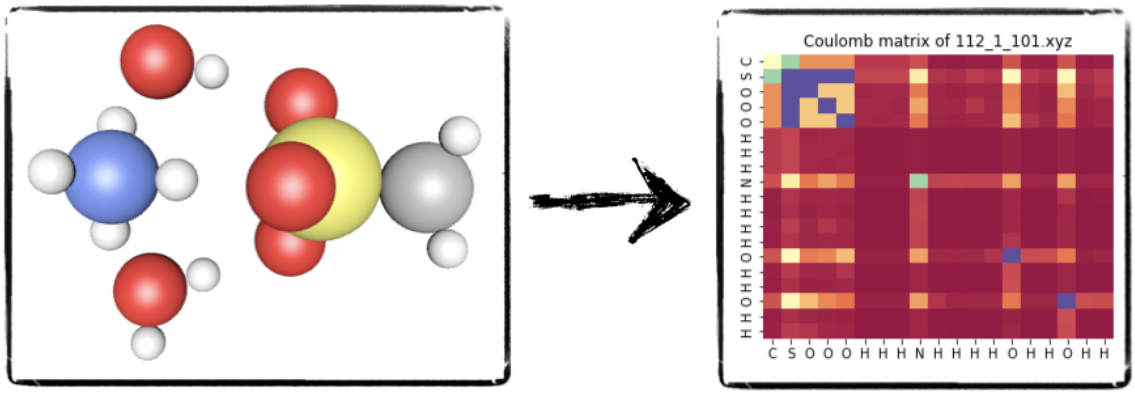


Figure 5.1: A molecular cluster of methanesulfonic acid, water and ammonia represented as an unordered Coulomb matrix. The values of each cell equals the Coulomb force between the atom pair.

5.2.2 Many-body Tensor Representation

Many-Body Tensor Representation [Huo and Rupp, 2017] represents atomic types and their relative positions with distributions collected into a multidimensional tensor. The first three terms that are most commonly used respectively correspond to atomic numbers (k_1), inverse distances between atoms in the molecule (k_2) and the cosines of angles between atoms (k_3). [Huo and Rupp, 2017, Himanen et al., 2020]

The values of k_1, k_2 and k_3 are broadened by using kernel density estimation with a gaussian kernel and summed over all combinations of atoms present in each equation. The broadening with gaussian functions makes it possible to represent any combination of atoms with MBTR's that have a constant size. The features used for ML algorithms come from the values of the functions $MBTR_1$, $MBTR_2$ and $MBTR_3$.

$$k_1(Z_i) = Z_i \tag{5.2}$$

$$k_2(\vec{R}_i, \vec{R}_j) = \frac{1}{|\vec{R}_i - \vec{R}_j|} \tag{5.3}$$

$$k_3(\vec{R}_i, \vec{R}_j, \vec{R}_k) = \cos\left(\angle(\vec{R}_i - \vec{R}_j, \vec{R}_k - \vec{R}_j)\right) \tag{5.4}$$

$$\mathrm{MBTR}_1^{Z_1}(x) = \sum_i^{|Z_1|} w_1^i \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-k_1(Z_i))^2}{2\sigma_1^2}} \tag{5.5}$$

$$\mathrm{MBTR}_2^{Z_1,Z_2}(x) = \sum_i^{|Z_1|}\sum_j^{|Z_2|} w_2^{i,j} \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{\left(x-k_2\left(R_i,R_j\right)\right)^2}{2\sigma_2^2}} \tag{5.6}$$

$$\mathrm{MBTR}_3^{Z_1,Z_2,Z_3}(x) = \sum_i^{|Z_1|}\sum_j^{|Z_2|}\sum_k^{|Z_3|} w_3^{i,j,k} \frac{1}{\sigma_3\sqrt{2\pi}} e^{-\frac{\left(x-k_3\left(R_i,R_j,R_k\right)\right)^2}{2\sigma_3^2}} \tag{5.7}$$

where $\vec{R}_i$ is the position vector of atom i, Z_i is the atomic number of atom i, σ is the standard deviation of the gaussian kernel and w is a weighting function that is used to control the importance of different terms. [Himanen et al., 2020]

The values of function $MBTR_2$ for the molecular cluster in this study are visualised in Fig. 5.2.

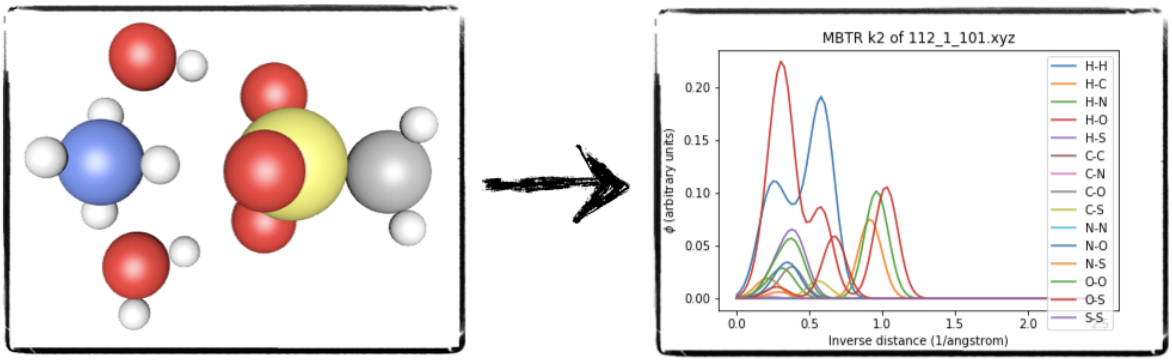


Figure 5.2: A molecular cluster of methanesulfonic acid, water and ammonia represented with Many-Body Tensor Representation. The figure shows the second term of the tensor ie. the inverse distances between each pair of atom types. The discreet peaks are gaussian smeared and hence the values correspond roughly to the amount of each distance present in the molecule.

5.2.3 Atom-Centered Symmetry Functions

Atom-Centered Symmetry Functions (ACSFs) consist of many-body symmetry functions that provide a descriptor for the chemical environment of each central atom i. Together the symmetry functions on all atoms represent the whole molecular structure. Typically, in case of a single chemical element, about 50 symmetry functions are used. A set of functions is obtained by defining them with different values for parameters ζ, λ, η, R_s and κ. The exact values of the sets of symmetry functions depend on the neighbouring atoms inside a chosen radial cut off R_c. The cut off function (5.8) defines the radius around each atom to take into account when creating the symmetry functions.[Behler, 2011]

The functions of ACSF descriptor are defined in Equations (5.9 - 5.13) and plotted for a molecular cluster of this study in Figure 5.3.

$$f_{\mathrm{c}}\left(R_{ij}\right) = \begin{cases} 0.5 \cdot \left[\cos\left(\frac{\pi R_{ij}}{R_c}\right) + 1\right] & \text{for } R_{ij} \leq R_c \\ 0 & \text{for } R_{ij} > R_c \end{cases} \tag{5.8}$$

where R_{ij} is the distance between atoms i and j, R_c is the cutoff radius, If R_{ij} is larger than R_c, the cut off function and its derivative become zero.

Radial functions The presence of neighboring atoms is detected by the radial two-body functions G_i^1, G_i^2 and G_i^3 (5.9,5.10,5.11) which consist of sums of two-body terms over all neighbours. They are physically related to effective coordination numbers: a set of radial symmetry functions can be considered as a description of the coordination at various distances from the central atom.

$$G_i^1 = \sum_j f_{\mathrm{c}}\left(R_{ij}\right) \tag{5.9}$$

$$G_i^2 = \sum_j \mathrm{e}^{-\eta\left(R_{ij}-R_{\mathrm{s}}\right)^2} \cdot f_{\mathrm{c}}\left(R_{ij}\right) \tag{5.10}$$

$$G_i^3 = \sum_j \cos\left(\kappa R_{ij}\right) \cdot f_{\mathrm{c}}\left(R_{ij}\right) \tag{5.11}$$

where j is the index of neighbouring atoms.

Function G^1 is simply the sum of the cut off functions with respect to j

Function G^2 is the sum of Gaussians multiplied by cut off functions. The width of the Gaussians is defined by a parameter η. The center of the Gaussians can be shifted to a certain radial distance by the parameter R_s. These "shifted" G^2 functions then are suitable to describe a spherical shell around the reference atom. For small values of η and $R_s = 0$ function G^2 reduces to function G^1. The radial distribution of neighbors

can be described by using a set of radial functions with different spatial extensions, e.g., G^1 functions with different cut off radii, or G^2 functions with different cut offs and/or η parameters.

Function G^3 Is a damped cosine function that should be used carefully and not without functions G^1 or G^2 because different G^3 functions might cancel out each other.

Angular functions The angular functions (equations 5.12, 5.13) are sums of three-body terms:

$$G_i^4 = 2^{1-\zeta} \sum_{j,k\neq i} (1 + \lambda \cos\theta_{ijk})^\zeta \cdot \mathrm{e}^{-\eta\left(R_{ij}^2+R_{ik}^2+R_{jk}^2\right)} \cdot f_\mathrm{c}\left(R_{ij}\right) \cdot f_\mathrm{c}\left(R_{ik}\right) \cdot f_\mathrm{c}\left(R_{jk}\right) \tag{5.12}$$

$$G_i^5 = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{all}} (1 + \lambda \cos\theta_{ijk})^\zeta \cdot \mathrm{e}^{-\eta\left(R_{ij}^2+R_{ik}^2\right)} \cdot f_\mathrm{c}\left(R_{ij}\right) \cdot f_\mathrm{c}\left(R_{ik}\right) \tag{5.13}$$

where $\theta_{ijk} = \arccos\left(\frac{\vec{R}_{ij}\cdot\vec{R}_{ik}}{R_{ij}\cdot R_{ik}}\right)$

The functions G_i^4 and G_i^5 have same angular part but different radial parts. The parameter λ can have the values $+1$ and -1 that shift the maxima. The angular resolution is provided by the parameter ζ: high ζ means narrower range of nonzero symmetry function values. A set of angular functions with different ζ-values can be used to obtain the distribution of angles centered at each reference atom similar to controlling the radial resolution of the radial functions $G2$ by parameter η. Further, this angular distribution can be determined at various distances from the central atom by a suitable choice of η and R_c, which control the radial part.

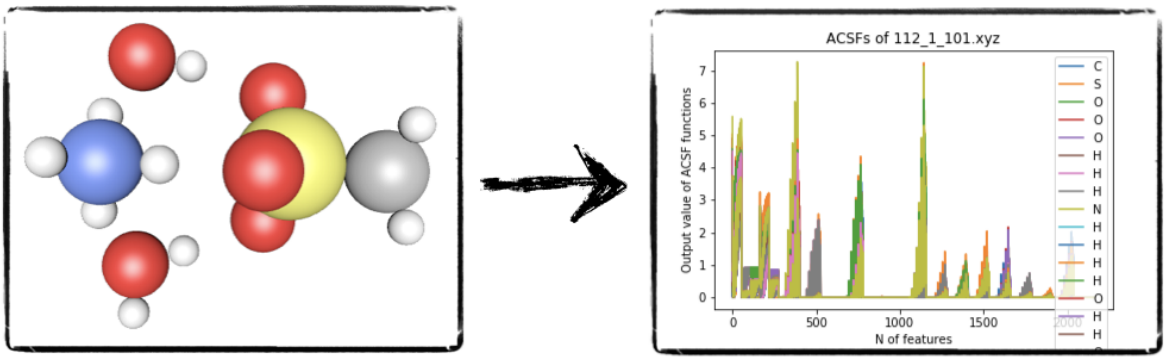


Figure 5.3: The Atom-Centered Symmetry Functions for a methanesulfonic acid, water and ammonia molecular cluster. The values of the functions come from the Equations 5.9 - 5.13 with multiple parameter values.

5.2.4 Smooth Overlap Atomic Postitions

The Smooth Overlap of Atomic Positions -descriptor encode local environment around an atom very accurately by integrating the overlap of smoothed out atomic positions and mapping them into coefficients of orthonormal basis functions. The smoothing is done by presenting the atoms' positions as gaussian functions: $\rho(r) = \sum_i e^{-(r-r_i)^2}$ as shown in Figure 5.4. SOAPs are calculated for all individual elements in the system and the values are concatenated in the end. Figure 5.5 shows the SOAP functions for a molecular cluster of this study.

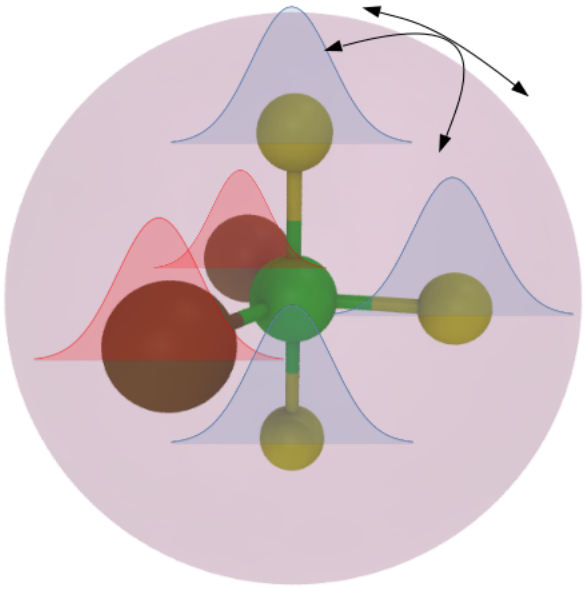


Figure 5.4: SOAP depicts the atom positions as gaussian functions. Image courtesy of Marc Jäger. Published with permission.

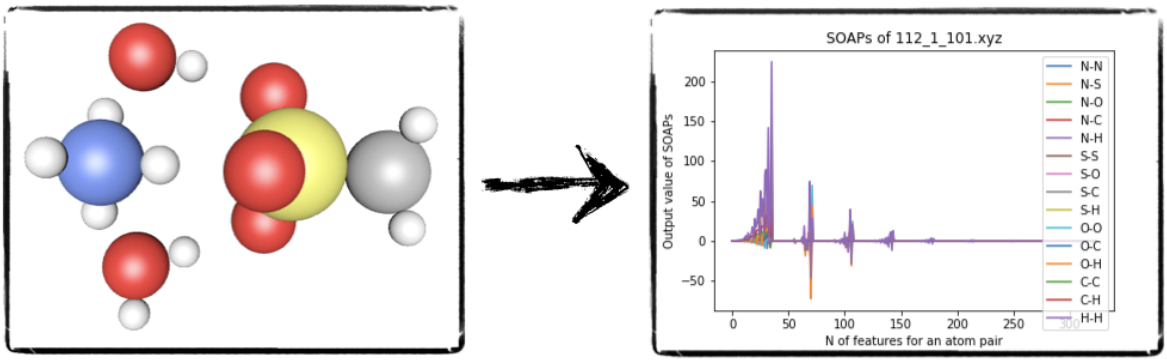


Figure 5.5: Density functions that make the Smooth Overlap of Atomic Positions for a of methanesulfonic acid, water and ammonia molecular cluster.

6. Comparison of the descriptors

Here the procedure of configurational sampling is described including parts of existing workflows and the new methods studied in this thesis. The main software and Python libraries used are: `JKCS` [Kubečka et al., 2019], `DScribe` [Himanen et al., 2020] and `ASE` [Hjorth Larsen et al., 2017] along with Python libraries for data wrangling eg. `KMeans` and `tsne` from `scikit-learn`, and `plotly` and `ASE` for visualizations. The Python version used was 3.7.3.

Figure 6.1 illustrates the basic steps of the proposed structure selection method. The structures are represented as descriptors and clustered with k-means algorithm. From the results a fraction of structures are selected.

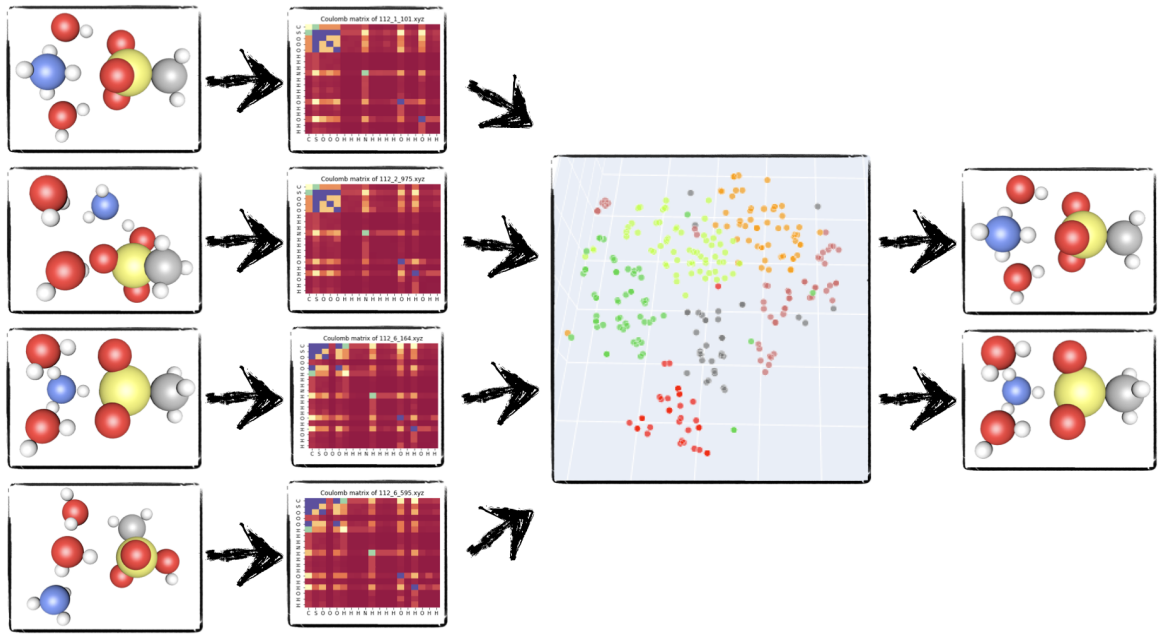
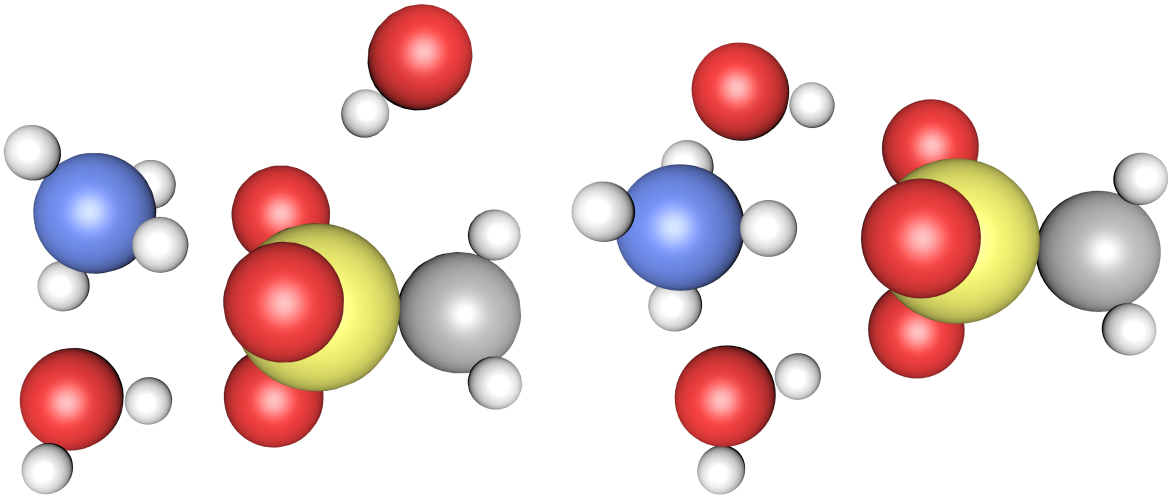


Figure 6.1: The proposed structure selection method illustrated. Molecular structures are represented with descriptors (here Coulomb matrices) and clustered. A fraction of structures is selected for further study.

6.1 The molecular cluster used in this study

The molecular cluster in this study consists of one methanesulfonic acid (MSA), one ammonia and two water molecules as shown in Figure 6.2. This cluster was chosen because the DFT and `XTB` data along with the structure coordinates for 577 configurations produced with `JKCS` were readily available. It is a small cluster with in total 4 molecules and 19 atoms of N, C, O, S, H. Small clusters have simple potential energy surfaces and hence necessary amount of configurations for configurational sampling is also small. The method studied here should be tested also with complex molecular cluster and a vast amount of different configurations.



(a) XTB-level optimised cluster. **(b)** DFT-level optimisation of the same cluster.

Figure 6.2: The molecular cluster in this study consists of one methanesulfonic acid (MSA), one ammonia and two water molecules. Here is one conformer of that molecular cluster first optimised with `XTB` and then further optimised with DFT. The geometry of the conformer changed drastically during DFT-optimisation. Colors: hydrogen - white, carbon - grey, nitrogen - blue, oxygen - red, sulfur - yellow.

6.2 JKCS

The basic usage of `JKCS` is described in the documentation found in the GitHub repository. The procedure starts with choosing the molecules that the studied clusters should consist of - here one methanesulfonic acid (MSA), ammonia and water as stated in the previous chapter. The script `JKCS0_copy` finds the predefined coordinates for the chosen molecules to be used by `ABCluster` [Zhang and Dolg, 2015, Zhang and Dolg, 2016] and produces an input.txt -file that is used to specify the amount of the chosen molecules in the cluster. Next `JKCS1_prepare` is used to make appropriate subfolders for the upcoming configurational sampling.

The first step in configurational sampling is to exhaustively explore and sample the whole potential energy surface ie. produce as many different conformations for the cluster as possible. This is done by `JKCS2_explore` that runs `ABCluster`. The parameters used for `JKCS2_explore` were `-pop 1000*M -gen 100 -lm 4000/NoC`: amount of initial guesses is 1000× the number of molecules (here 4), the bee colony population is 100 and the amount of structures saved is 4000/ number of combinations to form a cluster (here 2).

Running `ABCluster` results in molecular mechanics level energies and geometries for all saved structures. In order to gain fast, but better than MM-level approximation of energy and geometry `JKCS3_run` is used to run semi-empirical GFN-xTB calculations (Chapter 3.2.1) with `XTB` for all structures. While the `ABCluster` uses rigid molecules, now `XTB` relaxes the structures from their rigid MM-state.

The GFN-xTB-level results are collected with `JKCS4_collect`. The optimisation done by `XTB` might lead to structures that can be considered identical within a threshold and can be discarded. For example, when two configurations have differences less than 0.001 Hartree in energy and 0.1 Debye in dipole moment they are very likely to be identical. After the optimisation the structures are checked if they are still clusters: some structures may break apart as shown in Figure 6.3. They can be identified by having unusually large energy and the radius of gyration and can be discarded from the study.

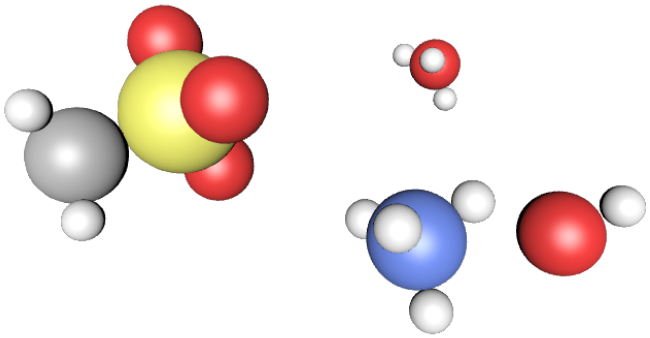


Figure 6.3: Some structures are "exploded" after `XTB` optimisation. They are identified from their unusually high energy and radius of gyration. This structure has the highest energy shown in Figure 6.6(a)

The usual configurational sampling procedure continues with `JKCS5_filter` that applies uniqueness selection by defining threshold values for energy, dipole and the radius of gyration and choosing only one from the groups of structures that have differences lower than the threshold values. The structure selection method presented in this study can be used to replace or rather to extend the functionality of `JKCS5_filter`. After the structure selection the DFT calculations are run with optimisations for remaining structures.

Resulting DFT structures and energies are used for further study along with the

corresponding XTB structures and energies. DFT calculations can also output Gibbs free energies (Eq. 3.8) which are preferred over electronic energies when studying of stabilities of molecular structures. The Figure 6.4 demonstrates the correlation between DFT electronic energies and Gibbs free energies and justifies the use of electronic energies when comparing DFT results with XTB electronic energies.

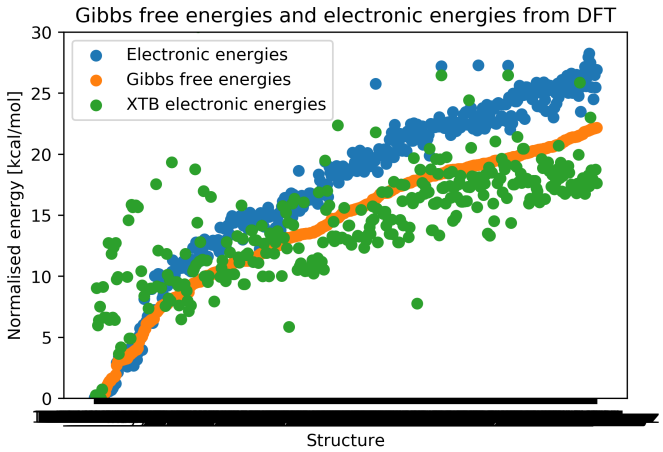


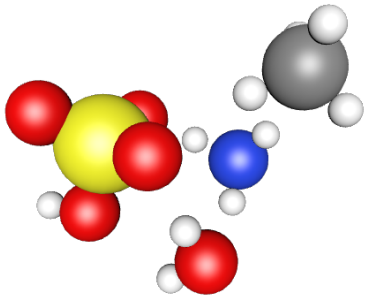
Figure 6.4: DFT Gibbs free energies and DFT electronic energies correlate, and the difference is considered so small that the electronic energies can be used when comparing with XTB energies.

Plotting energies of all structures against t-SNE components revealed an outlier in the dataset (see Fig. 6.5). It lies lower in the energy than the rest which would imply it corresponds to a global minimum, but a closer look revealed that it has different structure than the designed molecular cluster. During XTB optimisation one methanesulfonic acid had broken into methane and sulfuric acid. That does not happen in the nature spontaneously so the structure was discarded from the study.

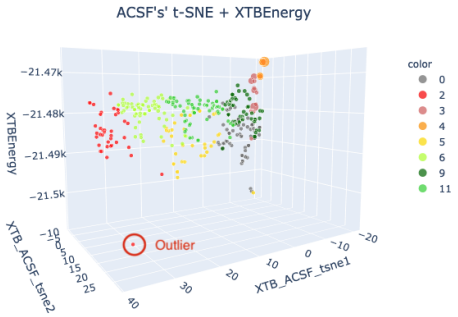
6.3 Creating descriptors

Energy, dipole and radius of gyration values from JKCS for all XTB and DFT structures are read into DataFrames of Python's pandas-library. The energies of XTB and DFT structures are normalised with respect to the lowest energy in the data set for both XTB and DFT respectively. The energy unit (originally Hartree) is converted to kcal/mol by multiplying with 627.509 kcal/mol. The coordinates of all structures are read with ASE in order to get a DScribe-compatible format. The structures are given as inputs for methods that create CM, MBTR, ACSF and SOAP descriptors of all the structures.

The hyperparameters for the descriptors are optimized in the original DScribe article [Himanen et al., 2020] with dataset from OQMD (http://oqmd.org) for predicting formation energies. The same parameters are used here with some adaptations.



(a) A molecular cluster of a sulfuric acid, water, ammonia and methane is different than original cluster of methanesulfonic acid, two waters and ammonia shown in Figure 6.2.



(b) The outlier is the lowest point in the graph

Figure 6.5: The structure with lowest energy corresponds to a molecular cluster with fragmented methanesulfonic acid. Such unintended and often undesired fragmentations can happen during the geometry optimisation calculations. The structure was discarded from the study.

Coulomb Matrices does not have actual hyperparameters. In order to account for independence of ordering mentioned in 5.2.1 the default setting for `permutation` is used.

MBTR hyperparameters are taken from the `DScribe` article supplementary material and adjusted according to the curves in Figure 5.2. Good values for the parameters result in peaks that are distinct but not too sharp. The broadness of the distribution is adjusted at each level k with parameter σ_k. Too small values lead to a delta-like distribution which would be too sensitive to differences in system configurations. The chosen values are $\sigma_1 = 0.6$, $\sigma_2 = 0.08$ and $\sigma_3 = 0.03$. The grid values are adjusted to take into account all data. For k_1 – the atomic numbers – reasonable values are up to `max`= 18 because the curve corresponding to sulphur with $N = 16$ reaches up to 18. For k_2 – the inverse distances – reasonable values are up to `max`= 2 because no atoms are closer to each other that 0.5 Å. For k_3 – the cosine of angles – the reasonable scale is $[-1, 1]$ because the output of cosine function is between -1 and 1. Values for scaling coefficients are 0.6 for k_2 and 0.2 for k_3. The cut off values are 0.001 for both k_2 and k_3.

The contributions of each k_1, k_2 and k_3 are normalized individually to unity by the euclidean length of each k-term. This is done by `DScribe` to ensure that the ML model considers the importance of each term equally. Otherwise the k-term with most features would dominate when k-means calculated the similarities between the structures. [Himanen et al., 2020]

ACSF hyperparameters are taken from the article by Jäger et al. [Jäger et al., 2018] and the cut off (R_{cut} is kept at 6.0 as in `DScribe` documentary.

SOAP hyperparameters are taken from the `DScribe` article supplementary material. Parameters for radial basis cutoff n_{max} and an angular basis (spherical harmonic) cutoff l_{max} are kept at 8. Larger values would introduce more computational cost by adding more features. For `sigma` the value is $\sigma = 0.4$.

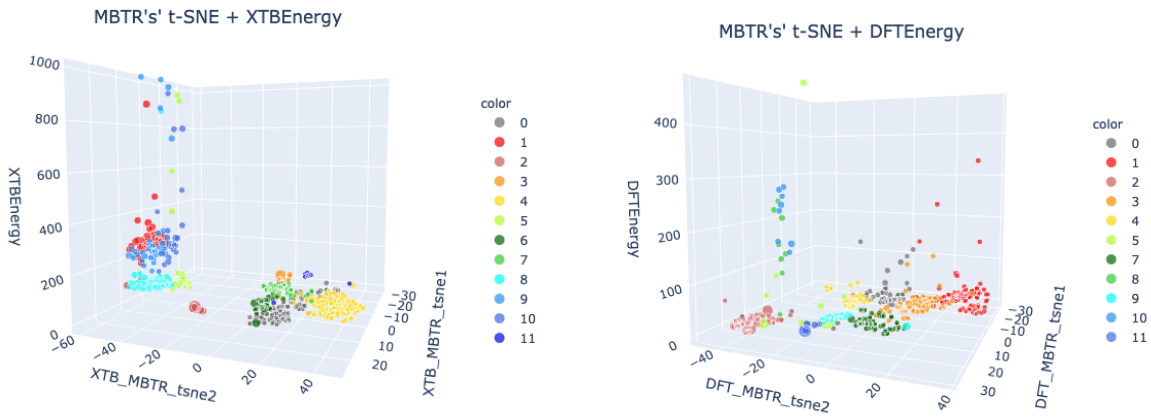
6.4 Clustering

The structures are clustered in XTB and DFT levels separately using all four descriptors individually yielding two times four groups of cluster labels. The clustering is done with `KMeans` from `sklearn`-library with the number of clusters $k = 12$ which is considered large enough to find possible different characteristics of the various molecular systems, but small enough so that for illustration purposes it is possible to find own distinctive color for each `kmeans` cluster. The clustering results for all descriptors are visualised in Figures like 6.6 to see how they map with the energies. Here the figure for MBTR is shown as an example but the other descriptors exhibit similar behaviour. Plotting MBTR's clustering results with energy as third axis shows how clustering results correspond to some extend with energy as well. The energy is not part of the data inputted to the k-means algorithm, but the energy values inside clusters are similar to each other. The extent to which that occurs depends on how well the features of a descriptor corresponds to the energy and on how many clusters (k) are initiated.

6.5 Structure selection

The structure selection method proposed here is based structure similarities determined by k-means from the values of descriptors. The values of the descriptors originate from the geometries of the molecules. There are two different strategies to use the information of cluster labels in a structure selection procedure: either pick a cluster that has on average the lowest energies and use all the structures in that cluster, or pick a random sample of structures from each cluster. The latter is preferable if the aim is to get as diverse selection of structures as possible into the next step of CS procedure. A hybrid version of the two strategies is to filter out the (k-means) clusters that have highest mean energies and pick a random sample from the remaining clusters.

With the last two structure selection methods the initial amount of k-means clusters k gives a way to adjust the amount of variety in the selected structures. With larger k and smaller size of the random sample taken from each cluster the structures will more likely be from different clusters and hence the variation between the structures is maximised.



(a) `XTB` structures labeled by `Kmeans` and their energies plotted on the z-axis.

(b) DFT structures labeled by `Kmeans` and their energies plotted on the z-axis.

Figure 6.6: The energies of `XTB` and DFT optimised structures are plotted on z-axis. The values are normalised to their respective minimas and converted to kcal/mol. The t-SNE components on x- and y- axis do not have a direct physical interpretation but they are used to project the clusters into two dimensions. Colours correspond to the cluster labels obtained from k-means algorithm. Especially on `XTB` level the structures that have significantly high energies are grouped together. Thus the descriptor used is able to represent the structures in a way that the values of the representation correspond to the energy.

6.6 Scoring the Descriptors

The structures are grouped by the cluster labels and mean energies for each cluster are calculated. The mean energies are used to select three of $k = 12$ clusters with the lowest mean energies and the names of structures belonging to the selected clusters are saved. The process is repeated for each descriptor. At this point the results are for each descriptor a list of `XTB` structures that belong to clusters withlowest mean `XTB`-energies and a list of DFT structures that belong to the clusters with the lowest mean DFT-energies. The DFT results are more accurate so they are considered as the ground truth that the `XTB` results can be compared to.

A score for each descriptor is calculated as a fraction of structures in `XTB` list found on DFT structure list:

$$\text{Score} = \frac{N(\texttt{XTB} \bigcap DFT)}{N(DFT)} \tag{6.1}$$

where $N(XTB \bigcap DFT)$ is the count of structures found in both lists and $N(DFT)$ is the count of structures in DFT list.

The results of k-means and therefore the cluster assignments depend on the random initial values for cluster centroids (see Chapter 6.4 for more detail) and the scoring for descriptors does show a dependency of the initial state of k-means algorithm. Thus the

lists of best mean energy structures can vary to some extent between two consecutive runs. Figure 6.7 illustrates the random behaviour as the scores ie. fractions of `XTB` structures in DFT list do not necessarily grow when the number of selected low energy clusters is changed and new clustering is performed.

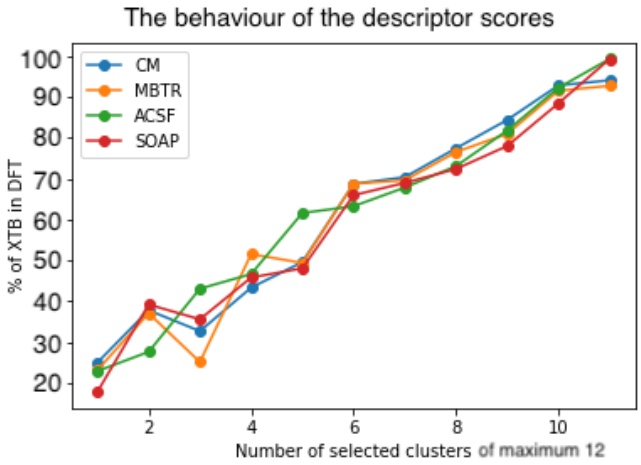


Figure 6.7: The scores for the descriptors as a function of the number of selected low energy clusters. The score is determined by comparing the lists of lowest energy structures for XTB and DFT results and calculating the percentage of XTB structures in the DFT structure list. The figure illustrates the stochastic behaviour of the scoring due to random initialisation of k-means clustering algorithm. In ideal situation the lines would go up linearly.

In order to make the scoring more reliable, the whole clustering process and scoring was repeated 500 times and each round scores for all descriptors were saved. The mean and standard deviation for the scores of each descriptor are calculated and plotted. The descriptor with best score is picked each round and a histogram of each descriptor occurring with the best score is calculated. The results are visualised in Figure 7.1. The process is identical for different descriptors except the change of the descriptor. Thus the score provides a measure of how rigorously each descriptor behaves with respect to the energies and gives insight to the choice of descriptor for stricture selection application.

6.7 Visualising the results and investigating cluster charasteristics

The results of k-means clustering done on each descriptor are visualised with `TSNE` from `sklearn`. The method is called t-Distributed Stochastic Neighbour Embedding (t-SNE) and it is a nonlinear dimensionality reduction method that shows if the clusters from k-means are separate or overlapping [van der Maaten and Hinton, 2008]. The results of t-SNE largely depend on the perplexity value [Wattenberg et al., 2016]. Thus a few different

perplexity values are tested (see Fig. A.5) and the default of 30 is used. Each descriptor outputs hundreds of thousands of features ie. *dimensions*. `TSNE` reduces the number of dimensions to an user specified value. In Figure 6.8 two dimensions are used and plotted against the DFT energy of each structure. The colors correspond to the cluster labels assigned by k-means algorithm and the size of the point corresponds to the radius of gyration of given structure. This plot is used to pick structures for visual inspection to verify if geometries of the structures exhibit features that correspond to their cluster assignments.

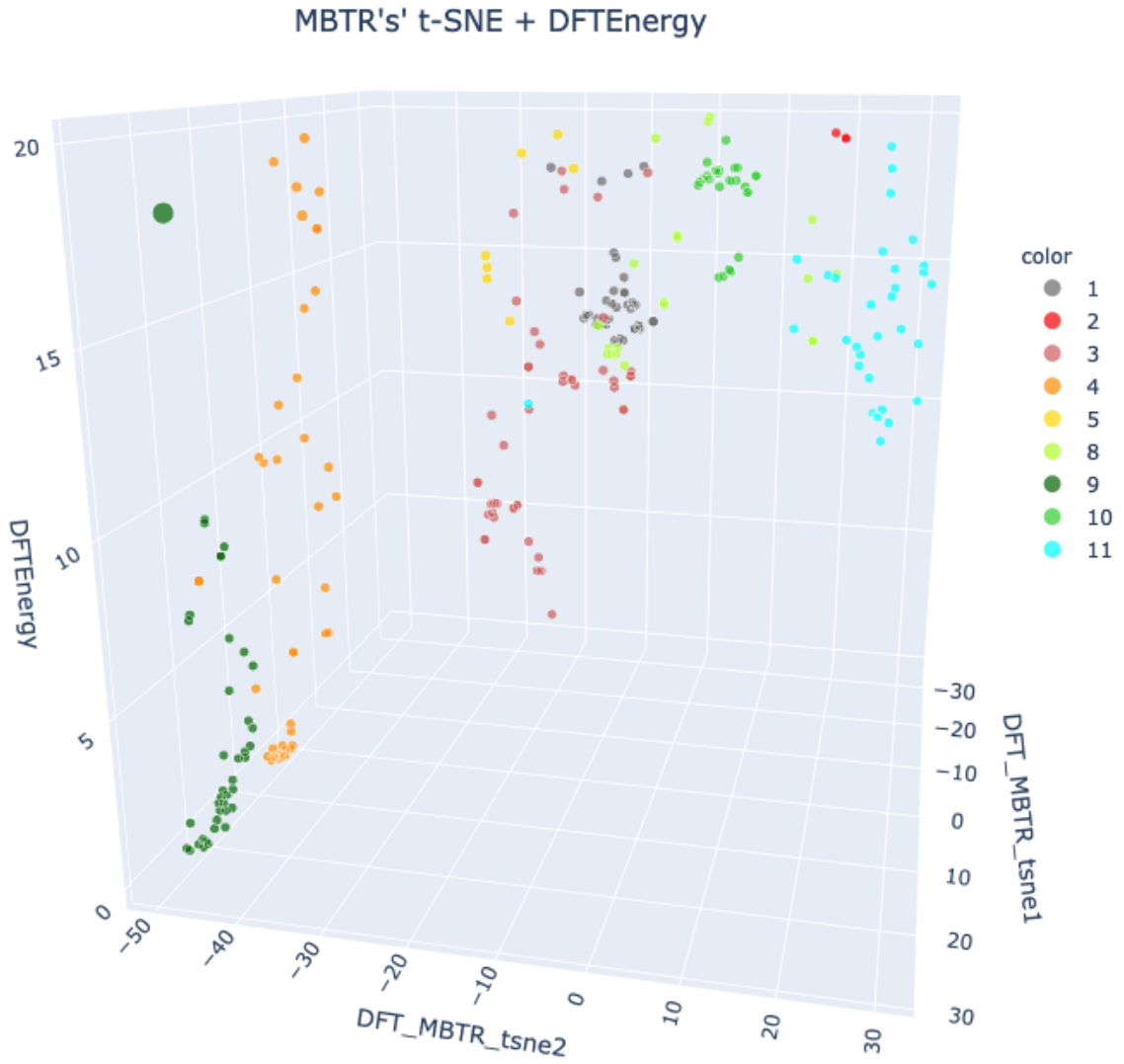


Figure 6.8: Clustered results of descriptors are plotted in two dimensions with t-SNE. The third dimension is the DFT energy of the structure normalised to zero and converted to kcal/mol. This plot is used to pick structures for visual inspection to verify if geometries of the structures exhibit features that correspond to their cluster assignments.

7. Results

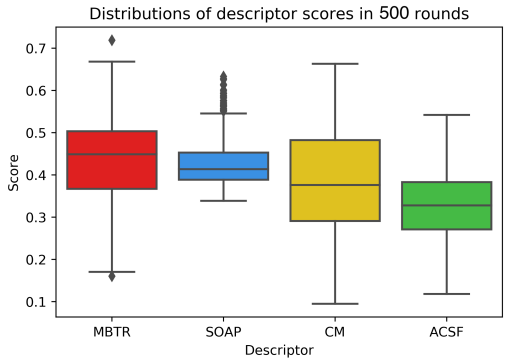
The k-means clustering algorithm is able to operate on features obtained from `DScribe` descriptors. The clustering results are plotted against the structure energies on both `XTB` and DFT level in Figures 6.6 and 6.8. The figures show that the structures with significantly higher energy are clustered together. As the energies are not part of the training data this implies that the descriptors can represent the structure in a way that the values of the descriptor to some extent correlate with the energy. Thus the hybrid structure selection method proposed in Chapter 6.5 can also be used for filtering out the structures with the highest energies.

7.1 Which descriptor to use for structure selection on atmospherical molecular clusters?

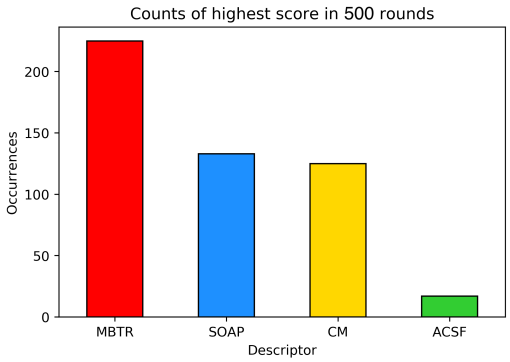
Figures 7.1(a) for the comparison of descriptors show that MBTR has the best scores of the four descriptors although SOAP has almost as good mean score. The deviations overlap with all the descriptors, but the histogram 7.1(b) shows that MBTR got the best score in almost half of the 500 test rounds. The results suggest the use of MBTR in the `JKCS` implementation of the structure selection.

7.2 Structure analysis for cluster features

Visualising the geometries of DFT optimised structures and their MBTR k_2 functions gives insight of how the clustering algorithm perceive the characteristics of the structures. The structures are selected by hand from k-means run on MBTR features and visualised by t-SNE as shown in Figure 6.8. Six structures with lowest DFT energy have really similar geometries and they are assigned into cluster number 9 shown in Fig. 6.8. The structures are visualised in Figure 7.2. Other geometries in cluster 9 show largely similar structures, but also some variety as shown in Figure A.1. The differences in MBTR k_2 values are modest but observable. The reason why the lowest three structures are in the same cluster is not clear and that could have been avoided if larger value of k were used.



(a) A boxplot showing mean and standard deviation for the descriptor scores after 500 rounds. MBTR has the highest mean, but SOAP has smallest standard deviation with many outliers above the mean.



(b) A histogram of occurrences for each descriptor with the best score. MBTR got the best score in almost half of the rounds.

Figure 7.1: 500 rounds of descriptor scores are saved and the mean, standard deviation and the histogram is shown. According to the figures MBTR is the best choice of descriptor in this application.

Comparison between two lowest distinct groups of structures (clusters 9 and 4 in Fig. 6.8) show that different hydrogen bonding pattern of water and ammonium molecules lead to the separation of the structures into different clusters. The upper three structures belong to the cluster 9 and they have all water and ammonia molecules hydrogen-bonded into the MSA. The lowes three structures belong to cluster 4 and they have a different hydrogen bonding pattern where one water is not hydrogen-bonded to the MSA but to the other water and ammonia instead.

The rotation of small molecules does not affect the clustering. The lowest energy structures of clusters 9 and 4 are visualised with their corresponding MBTR k_2 values in Figure A.2. The MBTR values reveal no obvious characteristics between the two clusters. Comparing structures from six different clusters reveal differences also in MBTR values as shown in Figure A.3. Moreover the figure shows that geometries with different protonation states of ammonia are clustered differently. More extensive analysis on geometries through various clusters indicate that at least MBTR enables k-means to differentiate between structures with ammonia and ammonium: clusters including both of ammonia and ammonium were not found. In Figure A.4 first three structures belong to cluster 1 and the rest belong to cluster 8. Both clusters 1 and 8 are on average higher in the energy and they contain only structures with ammonia. The features in MBTR that enable this distinction are not obvious to the human perception.

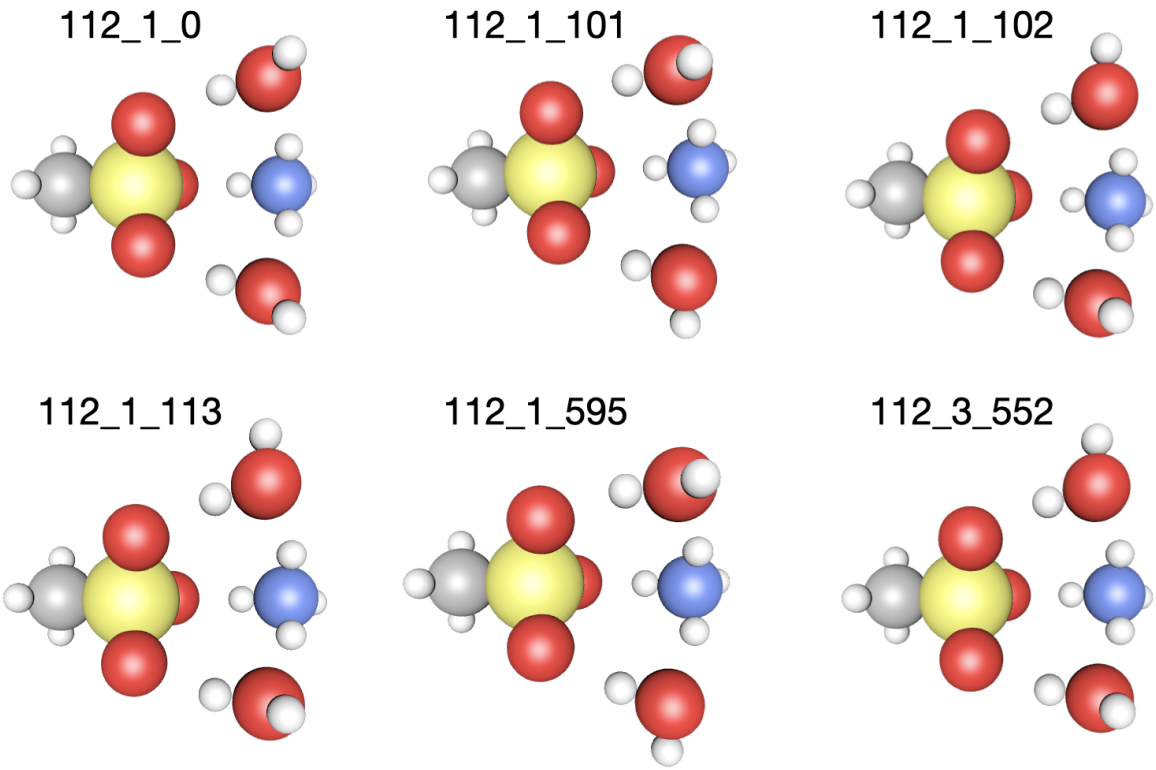


Figure 7.2: The DFT structures that correspond to the lowest DFT energies exhibit really similar geometries. Apparently eg. rotations of small molecules like water reflect such small-scale differences in MBTR that the structures are considered similar.

8. Conclusions

The objective of this study was to develop a new method for structure selection to be used in JKCS configurational sampling procedure and to gain further understanding about molecular representations. An essential step in the structure selection development was to determine which descriptor suits best for structure selection for atmospheric molecular clusters. A comparison between descriptors was made to guide the choice of descriptor for later use in JKCS. The comparison of the clusters was implemented with a custom scoring method because measuring the performance of the descriptors is not straightforward with the methods used in actual structure selection. The descriptors studied are in the DScribe-library and they were used to represent molecular clusters with Coulomb Matrices, Many-Body Tensors, Atom-Centered Symmetry functions and Smooth Overlaps of Atomic Positions. The structure selection was initiated by clustering the structures with k-means to group together structures that are similar to each other. A structure selection can be accomplished by choosing a subset of structures from clusters that have the lowest mean energies on XTB-level.

For further configural sampling research of atmospherically relevant molecular cluster structures the results of this study suggest the use of MBTR which proves the initial assumption correct. MBTR has by definition the best properties for studying energies and it has been demonstrated to function properly in other applications as well [Lumiaro, 2019]. If a second option is desired, averaged version of SOAP would be a good alternative according to the scores. It has been demonstrated to perform well in other applications as well [Jäger et al., 2018], but it should be kept in mind that it is originally designed to represent local properties whereas MBTR is designed as a global descriptor.

Moreover a visual inspection of structures from different clusters was conducted on MBTR results to gain further insight on the characteristics that the combination of MBTR and k-means can detect from the structures. Overall the visualisations shown in Figures A.3, A.2 and A.4 suggest that the rotations of a single small molecules like water and ammonia do not appear to have any effect on the MBTR values or clustering results. Then if those molecules change places or have different hydrogen bonding patterns the effect on clustering results can be detected. At least with MBTR k-means can distinguish structures with ammonium and ammonia: clusters including both of ammonia and ammonium were

not found which can be due to inability to scrutiny through all structures or to the fact that the results of k-means depend to some degree on the stochastic initialisation. An evident conclusion from the visual inspection is that with MBTR the intermolecular exchange of atoms such as a proton transfer reaction has more effect on the clustering results than rotational or translational changes in inside the molecular cluster.

In order to further validate or extend the research done for this study a few improvements and alternative methods are proposed. Any improvements should focus on these factors which affect the clustering results: the chosen descriptor and its hyperparameters, the chosen clustering method and its hyperparameters, the distance method used in clustering, the size of the dataset, and the ratio between the number of features in the descriptor and the size of the dataset.

For descriptors a detailed analysis of the descriptor hyperparameters is recommended. That could be conducted with brute force grid search or using an intelligent sampling of hyperparameter space like Bayesian optimization [Himanen et al., 2020]. In this study the hyperparameter spaces were not explored, but instead the values were taken from the literature. In the case of ACSF the hyperparameters were not necessarily optimised for small molecules which may partly explain the poor performance of ACSF. If a need rises more descriptors could be tested and even new descriptors can be engineered eg. by combining existing ones. It is difficult though to compare distances with multiple descriptors in parallel and hence some custom methods should me made also for that purpose. Some descriptors outside of `DScribe`-library are presented in references [Collins et al., 2017, Von Lilienfeld et al., 2015, Faber et al., 2018, Pronobis et al., 2018].

Clustering with k-means has many limitations which may affect the quality of the results also in this study. For example k-means does not learn the number of clusters k from the data and hence it has to be pre-defined when the algorithm starts. [Raykov et al., 2016, James et al., 2017] With new dataset produced by `JKCS` there is no prior knowledge of how many distinguishable groups of structures the dataset contains and hence the structure selection method would benefit from a clustering method that is able to learn this factor from the data. On this study it appears that $k = 12$ is not large enough to capture all the different characteristics of the structures into separate clusters. For example the clusters 9 and 4 in Figure 6.8 could have been splitted at approximately 5 kcal/mol.

Furthermore the shape of the data can affect the clustering results as k-means algorithm works well if the data has clusters with spherical shape. The algorithm of k-means always tries to construct a spherical clusters around the centroids. Thus if the clusters have more complicated geometric shapes, k-means may not respect the shape of the data

and instead tries allocate data into spherical clusters. [Raykov et al., 2016]

With molecular structure data the shapes of clusters are not well-defined but rather obscure and varying in density. High cluster density indicate many datapoints that have high similarity and low cluster density imply fewer datapoints of lower similarity. That would suggest the use of methods that can adapt to different cluster densities. For example DBSCAN and OPTICS are density-based clustering methods that could both learn the amount of clusters k from the data and identify if the structures form some dense clusters that are not spherical [Oskolkov, 2019]. DBSCAN and OPTICS also can leave some datapoints unlabeled and therefore detect outliers which would be beneficial in structure selection applications.

Dataset size and the number of features affect the clustering results. In general the results of machine learning methods are more accurate with larger datasets. The important factor is the ratio between the dataset size and the number of features. It is suggested that the minimum dataset size should be $70 \times m$ where m is the number of features [Dolnicar et al., 2014] or 2^m [Formann, 1984]. Both sources imply that also k-means should have dataset size a couple orders of magnitude larger than the number of features. In this study the numbers of features were 361,9500,2165,4860 for CM, MBTR, ACSF and SOAP respectively and the data set size was 576* structures. The numbers of features are especially high with MBTR and SOAP but as they already got the best scores the conclusion is that they could perform even better when the dataset size is 10 000 or more. In order to reduce the number of features some dimensionality reduction techniques like Principal Component Analysis (PCA) [Jollife and Cadima, 2016] or Independent Component Analysis (ICA) [Hyvärinen and Oja, 2000] should be tested. PCA has been used with descriptors before [Jäger et al., 2018].

The distance metrics is the method of how the similarity between two datapoints is measured. The Euclidian distance which was used in this study is not the most favourable choice with the descriptors because of the large number of features that the descriptors output. [Oskolkov, 2019]. Euclidian distance is used with descriptors in other studies [Huo and Rupp, 2017] but some other metrics like Manhattan distance ("a taxi cab -distance"), cosine distance or Tanimoto coefficient should be tested. [Bora and Gupta, 2014, Irani et al., 2016]

Other studies would include applying these methods on various molecular systems to test the applicability of the methods presented in this study. There is no technical issues to restrain the type of molecular structure that these methods can be applied to.

*Originally there were 577 structures but the fragmented outlier was discarded.

Any field of computational molecular structure research like drug design could eventually benefit of procedures described in this study. Alternative further studies in atmospheric molecular cluster research would include for example testing RuNNer which is a neural network developed for potential energy surface calculations [Behler, 2018], investigating the possibilities of transfer learning which could be used for mapping the lower energy level datapoints into higher energy level [Smith et al., 2019b], or Gaussian Approximation Potentials that could be used for fitting interatomic potentials based on quantum chemistry calculations [Bartõk and Csányi, 2015].

Bibliography

[Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 07-09-January-2007:1027–1035.

[Bartõk and Csányi, 2015] Bartõk, A. P. and Csányi, G. (2015). Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry*, 115(16):1051–1057.

[Bartók et al., 2013] Bartók, A. P., Kondor, R., and Csányi, G. (2013). On representing chemical environments. *Physical Review B - Condensed Matter and Materials Physics*, 87(18):1–16.

[Behler, 2011] Behler, J. (2011). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *Journal of Chemical Physics*, 134(7).

[Behler, 2018] Behler, J. (2018). Runner-a neural network code for high-dimensional potential-energy surfaces. *Universität Göttingen*.

[Besel, 2020] Besel, V. (2020). Impact of Quantum Chemistry Parameters and Model Settings on Predicted Atmospheric Particle Formation. Master's thesis, University of Helsinki.

[Bora and Gupta, 2014] Bora, M. D. J. and Gupta, D. A. K. (2014). Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab. 5(2):2501–2506.

[Born and Oppenheimer, 1927] Born, M. and Oppenheimer, R. (1927). Zur quantentheorie der molekeln. *Ann. Phys.*, 389:457–484.

[Brooks et al., 1983] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217.

[Case et al., 2010] Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossvary, I., Wong, K. F., Paesani, F., Vanicek, J., Liu, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M.-J., Hornak, V., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., Kollman, P. A., and Roberts, B. P. (2010). Amber 11.

[Collins et al., 2017] Collins, C. R., Gordon, G. J., von Lilienfeld, O. A., and Yaron, D. J. (2017). Constant Size Molecular Descriptors For Use With Machine Learning.

[De et al., 2016] De, S., Bartók, A. P., Csányi, G., and Ceriotti, M. (2016). Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769.

[Dolnicar et al., 2014] Dolnicar, S., Grün, B., Leisch, F., and Schmidt, K. (2014). Required Sample Sizes for Data-Driven Market Segmentation Analyses in Tourism. *Journal of Travel Research*, 53(3):296–306.

[Dua and Graff, 2017] Dua, D. and Graff, C. (2017). UCI machine learning repository.

[Elm, 2019] Elm, J. (2019). An atmospheric cluster database consisting of sulfuric acid, bases, organics, and water. *ACS Omega*, 4(6):10965–10974.

[Faber et al., 2018] Faber, F. A., Christensen, A. S., Huang, B., and Von Lilienfeld, O. A. (2018). Alchemical and structural distribution based representation for universal quantum machine learning. *Journal of Chemical Physics*, 148(24).

[Fock, 1930] Fock, V. (1930). Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Z. Angew. Phys.*, 61(1):126–148.

[Formann, 1984] Formann, A. K. (1984). Die latent-class-analyse: Einführung in die theorie und anwendung. beltz.

[Gebauer et al., 2018] Gebauer, N. W. A., Gastegger, M., and Schütt, K. T. (2018). Generating equilibrium molecules with deep neural networks.

[Ghosh et al., 2019] Ghosh, K., Stuke, A., Todorović, M., Jørgensen, P. B., Schmidt, M. N., Vehtari, A., and Rinke, P. (2019). Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra. *Advanced Science*, 6(9).

[Grimme et al., 2017] Grimme, S., Bannwarth, C., and Shushkov, P. (2017). A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational

Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *Journal of Chemical Theory and Computation*, 13(5):1989–2009.

[Hansen et al., 2015] Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., Von Lilienfeld, O. A., Müller, K. R., and Tkatchenko, A. (2015). Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *Journal of Physical Chemistry Letters*, 6(12):2326–2331.

[Hartree, 1928] Hartree, D. R. (1928). The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. *Math. Proc. Cambridge Philos. Soc.*, 24(1):89–110.

[Heidenreich, 2018] Heidenreich, H. (2018). The Future with Reinforcement Learning.

[Himanen et al., 2020] Himanen, L., Jäger, M. O., Morooka, E. V., Federici Canova, F., Ranawat, Y. S., Gao, D. Z., Rinke, P., and Foster, A. S. (2020). DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949.

[Hjorth Larsen et al., 2017] Hjorth Larsen, A., JØrgen Mortensen, J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., Hermes, E. D., Jennings, P. C., Bjerre Jensen, P., Kermode, J., Kitchin, J. R., Leonhard Kolsbjerg, E., Kubal, J., Kaasbjerg, K., Lysgaard, S., Bergmann Maronsson, J., Maxson, T., Olsen, T., Pastewka, L., Peterson, A., Rostgaard, C., SchiØtz, J., Schütt, O., Strange, M., Thygesen, K. S., Vegge, T., Vilhelmsen, L., Walter, M., Zeng, Z., and Jacobsen, K. W. (2017). The atomic simulation environment - A Python library for working with atoms. *Journal of Physics Condensed Matter*, 29(27).

[Hohenberg and Kohn, 1964] Hohenberg, P. and Kohn, W. (1964). Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871.

[Holliday et al., 2004] Holliday, J. D., Rodgers, S. L., Willett, P., Chen, M. Y., Mahfouf, M., Lawson, K., and Mullier, G. (2004). Clustering files of chemical structures using the fuzzy k-means clustering method. *Journal of Chemical Information and Computer Sciences*, 44(3):894–902.

[Huo and Rupp, 2017] Huo, H. and Rupp, M. (2017). Unified Representation of Molecules and Crystals for Machine Learning. (July).

[Hyvärinen and Oja, 2000] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

[Irani et al., 2016] Irani, J., Pise, N., and Phatak, M. (2016). Clustering Techniques and the Similarity Measures used in Clustering: A Survey. *International Journal of Computer Applications*, 134(7).

[Jäger et al., 2018] Jäger, M. O., Morooka, E. V., Federici Canova, F., Himanen, L., and Foster, A. S. (2018). Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials*, 4(1).

[James et al., 2017] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017). *An Introduction to Statistical Learning*. Springer New York Heidelberg Dordrecht London.

[Jensen, 2017] Jensen, F. (2017). *Introduction to Computational Chemistry*. John Wiley & Sons, Inc.

[Jollife and Cadima, 2016] Jollife, I. T. and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065).

[Jordan and Mitchell, 2015] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. 349(6245).

[Jung et al., 2020] Jung, H., Stocker, S., Kunkel, C., Oberhofer, H., Han, B., Reuter, K., and Margraf, J. T. (2020). Size-Extensive Molecular Machine Learning with Global Representations. *ChemSystemsChem*, 1900052:1–8.

[Kerminen et al., 2012] Kerminen, V. M., Paramonov, M., Anttila, T., Riipinen, I., Fountoukis, C., Korhonen, H., Asmi, E., Laakso, L., Lihavainen, H., Swietlicki, E., Svenningsson, B., Asmi, A., Pandis, S. N., Kulmala, M., and Petäjä, T. (2012). Cloud condensation nuclei production associated with atmospheric nucleation: A synthesis based on existing literature and new results. *Atmospheric Chemistry and Physics*, 12(24):12037–12059.

[Koch and Holthausen, 2001] Koch, W. and Holthausen, M. C. (2001). *A Chemist's Guide to Density Functional Theory*. Wiley-VCH Verlag GmbH, second edition edition.

[Kohn and Sham, 1965] Kohn, W. and Sham, L. J. (1965). Self-Consistent Equations Including Exchange and Correlation Effects. 140(4A).

[Kubečka et al., 2019] Kubečka, J., Besel, V., Kurtén, T., Myllys, N., and Vehkamäki, H. (2019). Configurational Sampling of Noncovalent (Atmospheric) Molecular Clusters: Sulfuric Acid and Guanidine. *Journal of Physical Chemistry A*, 123(28):6022–6033.

[Lentze, 2015] Lentze, G. (2015). Newsletter No. 143 - Spring 2015. (143).

[Lo et al., 2018] Lo, Y. C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1538–1546.

[Lumiaro, 2019] Lumiaro, E. (2019). *Predicting Gas-Particle Partitioning Properties of Atmospheric Molecules Using Kernel Ridge Regression*. PhD thesis, Aalto University.

[McGrath et al., 2012] McGrath, M. J., Olenius, T., Ortega, I. K., Loukonen, V., Paasonen, P., Kurtén, T., Kulmala, M., and Vehkamäki, H. (2012). Atmospheric Cluster Dynamics Code: A flexible method for solution of the birth-death equations. *Atmospheric Chemistry and Physics*, 12(5):2345–2355.

[Montavon et al., 2015] Montavon, G., Hansen, K., Fazli, S., Rupp, M., Biegler, F., Ziehe, A., Tkatchenko, A., Von Lilienfeld, O. A., and Müller, K. R. (2015). Learning Invariant Representations ofMolecules for Atomization Energy Prediction. *Journal of Physical Chemistry Letters*, 6(12):2326–2331.

[Myhre et al., 2013] Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., and Zhang, H. (2013). *Anthropogenic and natural radiative forcing*, pages 659–740. Cambridge University Press, Cambridge, UK.

[Oskolkov, 2019] Oskolkov, N. (2019). How to cluster in High Dimensins.

[Partanen et al., 2016] Partanen, L., Vehkamaki, H., Hansen, K., Elm, J., Henschel, H., Kurten, T., Halonen, R., and Zapadinsky, E. (2016). Effect of conformers on free energies of atmospheric complexes. *Journal of Physical Chemistry A*, 120(43):8613–8624.

[Perdew and Schmidt, 2001] Perdew, J. P. and Schmidt, K. (2001). Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings*, 577(1):1–20.

[Pronobis et al., 2018] Pronobis, W., Tkatchenko, A., and Müller, K. R. (2018). Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules. *Journal of Chemical Theory and Computation*, 14(6):2991–3003.

[Ramakrishnan et al., 2015] Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. (2015). Big data meets quantum chemistry approximations: The Δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096.

[Raykov et al., 2016] Raykov, Y. P., Boukouvalas, A., Baig, F., and Little, M. A. (2016). What to do when K-means clustering fails: A simple yet principled alternative algorithm. *PLoS ONE*, 11(9):1–28.

[Rupp, 2015] Rupp, M. (2015). Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry*, 115(16):1058–1073.

[Rupp et al., 2012] Rupp, M., Tkatchenko, A., Müller, K. R., and Von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):1–5.

[Schütt et al., 2019a] Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K. R., and Maurer, R. J. (2019a). Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10(1):1–10.

[Schütt et al., 2019b] Schütt, K. T., Kessel, P., Gastegger, M., Nicoli, K. A., Tkatchenko, A., and Müller, K. R. (2019b). SchNetPack: A Deep Learning Toolbox for Atomistic Systems. *Journal of Chemical Theory and Computation*, 15(1):448–455.

[Seko et al., 2017] Seko, A., Hayashi, H., Nakayama, K., Takahashi, A., and Tanaka, I. (2017). Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14):1–11.

[Smith et al., 2019a] Smith, J. S., Nebgen, B. T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O., and Roitberg, A. E. (2019a). Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications*, 10(1).

[Smith et al., 2019b] Smith, J. S., Nebgen, B. T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O., and Roitberg, A. E. (2019b). Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications*, 10(1):1–8.

[Stuke et al., 2019] Stuke, A., Todorović, M., Rupp, M., Kunkel, C., Ghosh, K., Himanen, L., and Rinke, P. (2019). Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *Journal of Chemical Physics*, 150(20).

[Svensmark et al., 2017] Svensmark, H., Enghoff, M. B., Shaviv, N. J., and Svensmark, J. (2017). Increased ionization supports growth of aerosols into cloud condensation nuclei. *Nature Communications*, 8(1):1–9.

[van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE Laurens. *Journal of Machine Learninf Research*, 9:2579–2605.

[Von Lilienfeld et al., 2015] Von Lilienfeld, O. A., Ramakrishnan, R., Rupp, M., and Knoll, A. (2015). Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry*, 115(16):1084–1093.

[Wattenberg et al., 2016] Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill*.

[Wei Zhong et al., 2005] Wei Zhong, Altun, G., Harrison, R., Tai, P. C., and Yi Pan (2005). Improved k-means clustering algorithm for exploring local protein sequence motifs representing common structural property. *IEEE Transactions on NanoBioscience*, 4(3):255–265.

[Yao et al., 2018] Yao, K., Herr, J. E., Toth, D. W., McKintyre, R., and Parkhill, J. (2018). The TensorMol-0.1 model chemistry: A neural network augmented with long-range physics. *Chemical Science*, 9(8):2261–2269.

[Zhang and Dolg, 2015] Zhang, J. and Dolg, M. (2015). ABCluster: The artificial bee colony algorithm for cluster global optimization. *Physical Chemistry Chemical Physics*, 17(37):24173–24181.

[Zhang and Dolg, 2016] Zhang, J. and Dolg, M. (2016). Global optimization of clusters of rigid molecules using the artificial bee colony algorithm. *Physical Chemistry Chemical Physics*, 18(4):3003–3010.

Appendix A. Attachments

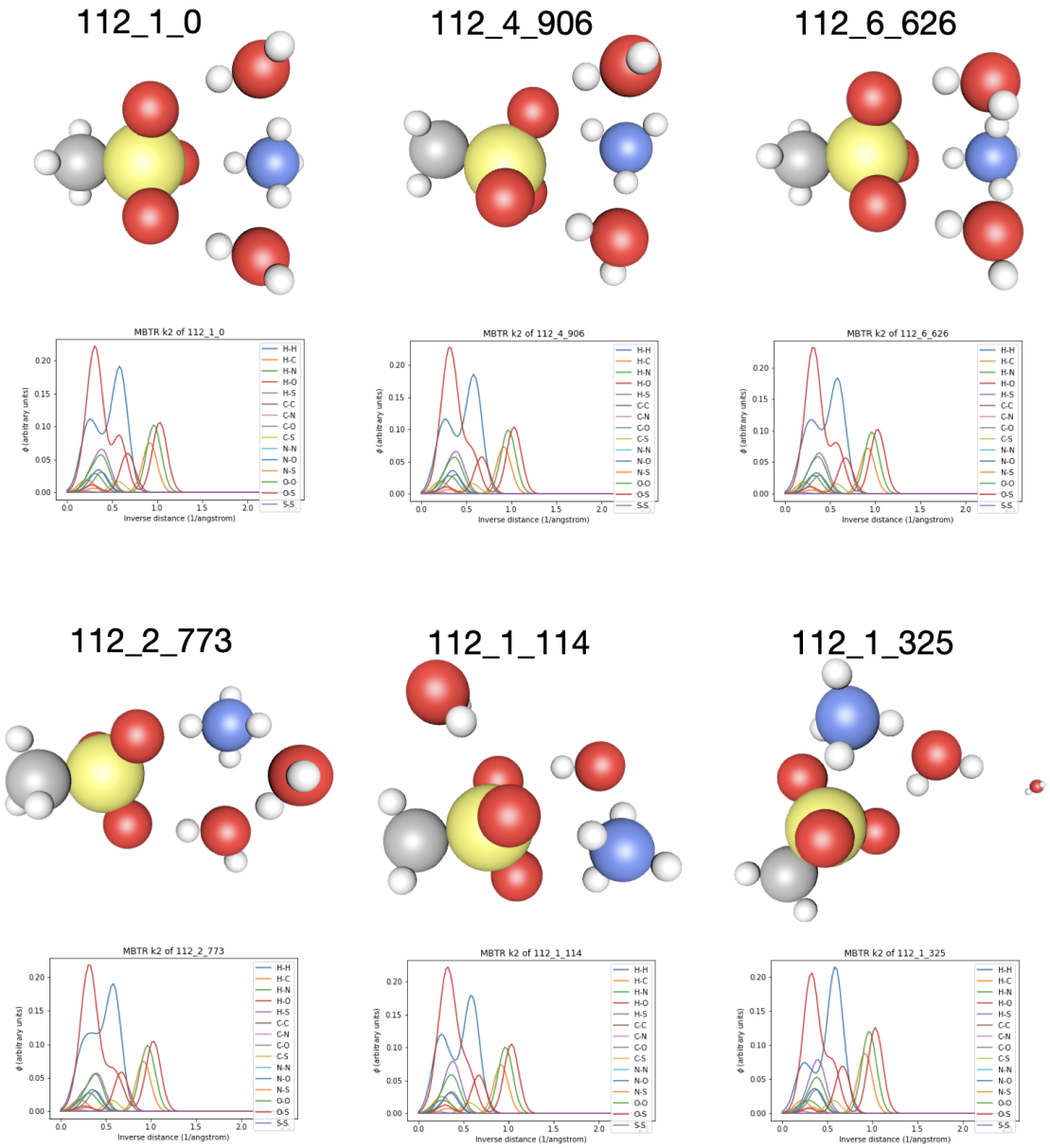


Figure A.1: The cluster which includes the structures with lowest DFT energies has some consistency, but also some variety among structure geometries. Structure 112_1_325 has lost one water molecule but the difference in MBTR is small because the x-axis is inverted and hence large differences in large distances make small difference in MBTR values.

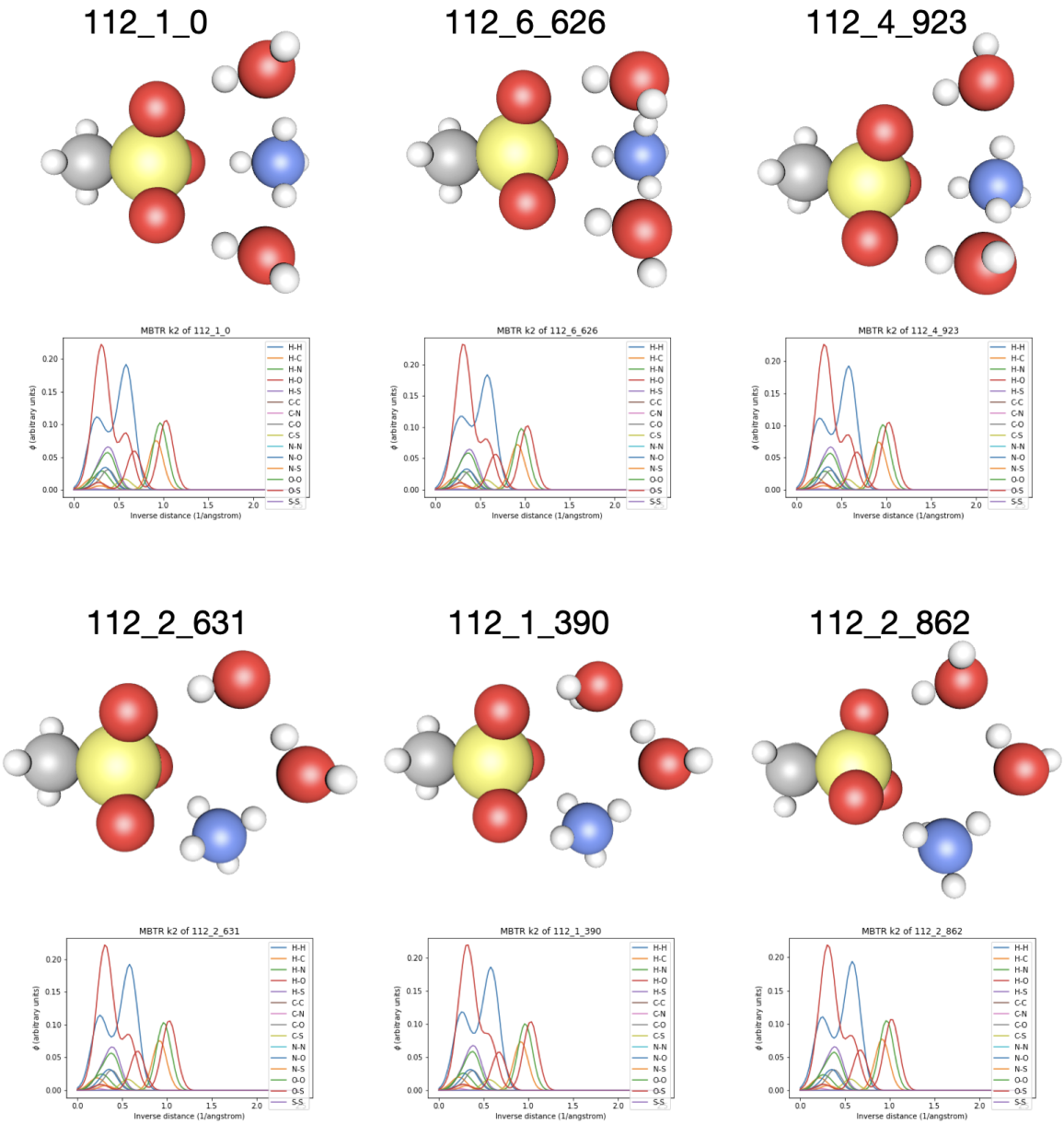


Figure A.2: Structures on the upper row belong to the cluster with lowest energies and the structures on the lower row to the cluster with second lowest energies. Difference between lowest energy structures of two these two "best" clusters is obvious: the order of water and ammonium molecules changes - but the differences in MBTR are subtle.

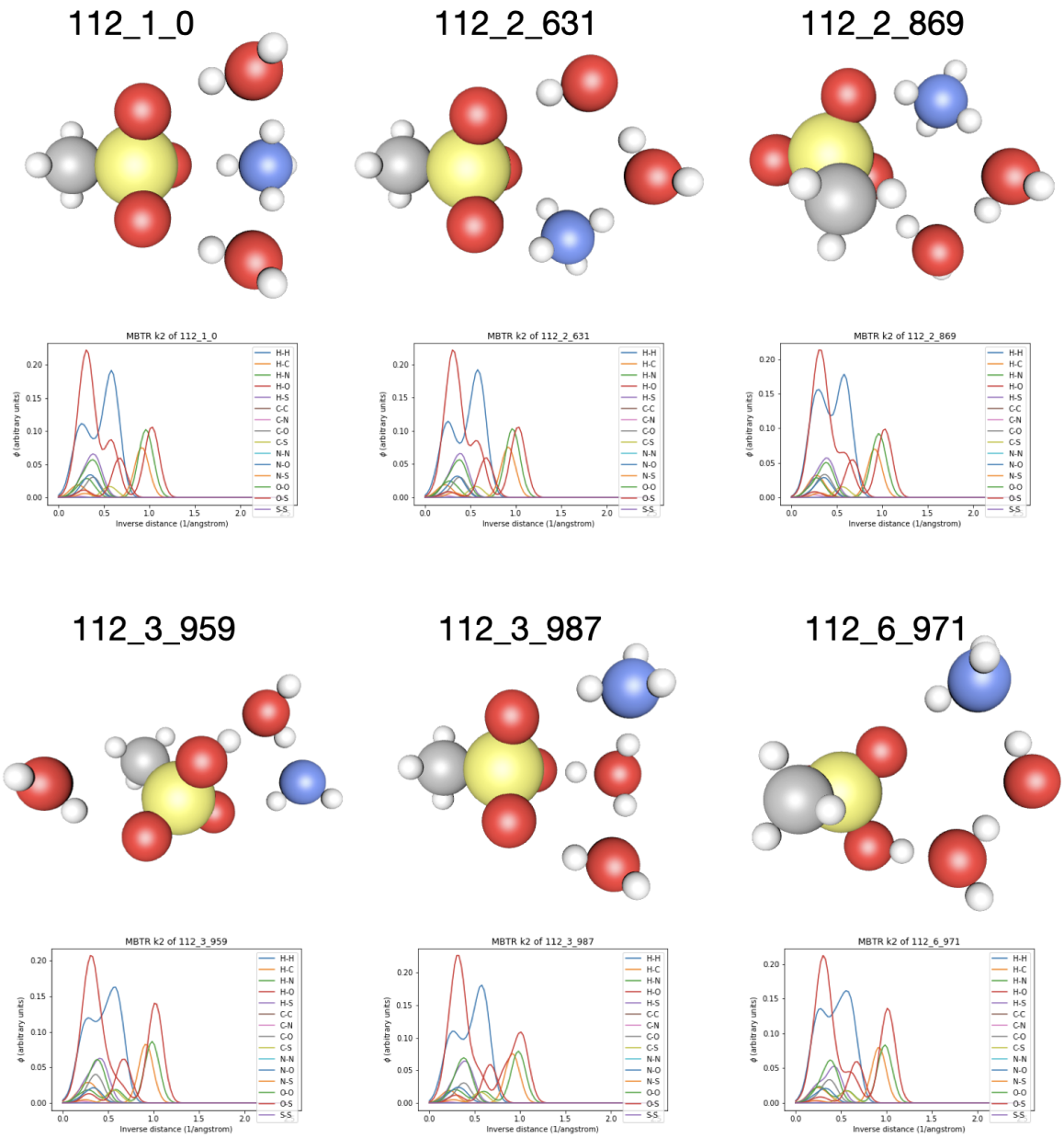


Figure A.3: The lowest energy structures of six different clusters demostrate how different geometries are grouped into separate clusters. Moreover structures in different clusters have distinct protonation states of ammonia. The differences in MBTR concerning proton transfer are not apparent to human perception.

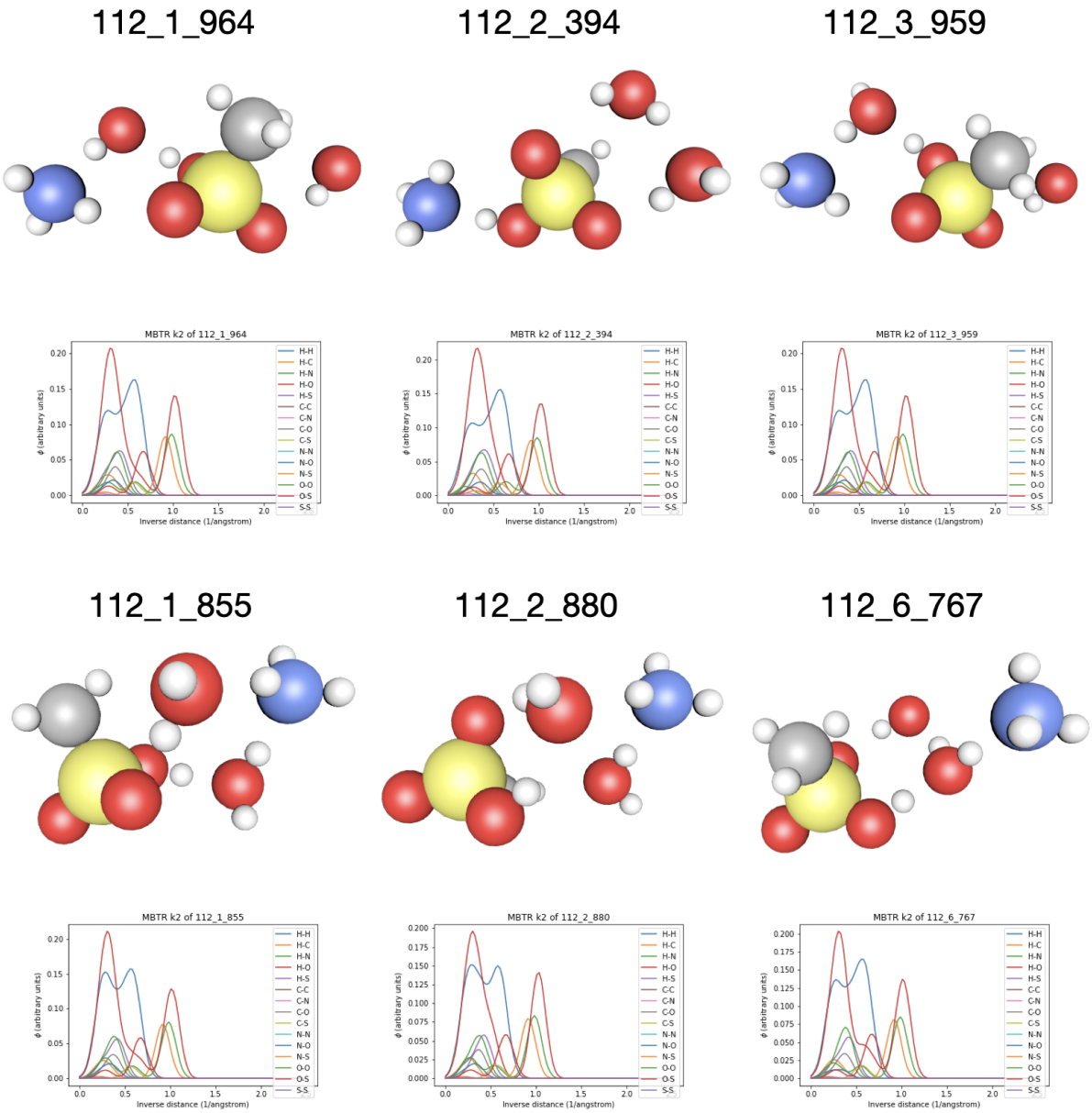


Figure A.4: Structures on the upper row belong to one cluster and the structures on the lower row to another. Both clusters only include structures that have ammonia with three protons.

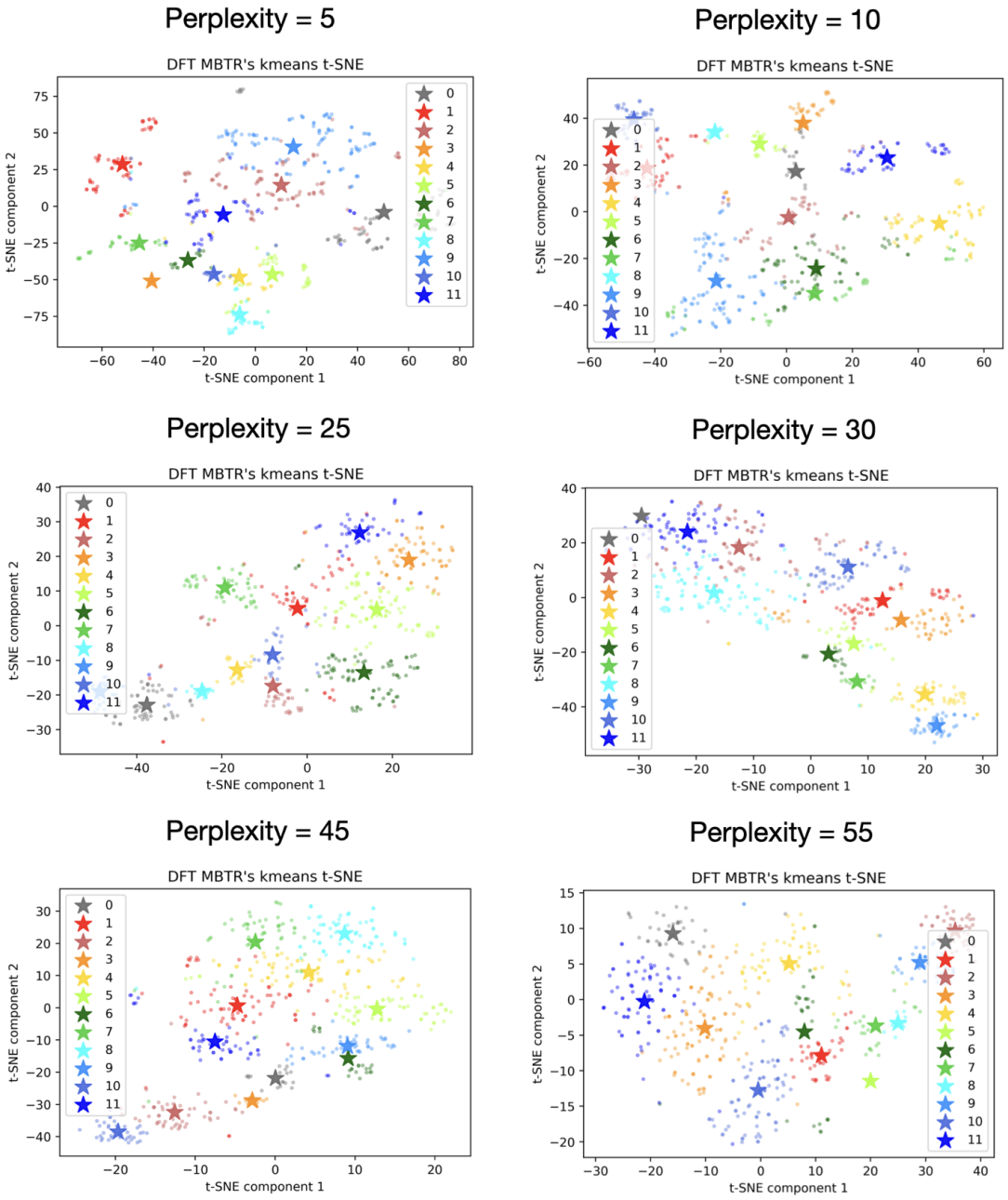


Figure A.5: Different perplexity values for t-SNE visualisation of clusters. According to the figures the perplexity value does not make a huge difference - good choices being for example 30 or 55. The recommended values are between 5-50 [Wattenberg et al., 2016] and hence the default 30 were used.